SEM uff 12 EST

12ª SEMANA DA ESTATÍSTICA

www.semest.uff.br

De 18 a 20 de outubro de 2021

Universidade Federal Fluminense

AGENDA ACADÊMICA

Departamento de Estatística uff

Coordenação do Curso de Graduação em Estatística uff

LES uff Universidade Federal Fluminense

12ª Semana da Estatística

18-20 de Outubro de 2021

UFF, Niterói, Rio de janeiro, Brasil

Anais do Evento

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade Federal Fluminense

ISBN: 978-65-00-36246-6

Sumário

Sobre

12ª Semana da Estatística

A Semana da Estatística (SEMEST) é um evento que ocorre dentro da Agenda Acadêmica da Universidade Federal Fluminense (UFF). Tradicionalmente o evento conta com palestras e minicursos, abordando diferentes áreas de aplicação da Estatística, além de sessões com a apresentação de trabalhos submetidos.

Em 2021, em sua primeira edição completamente remota/online, o evento contou com palestrantes de diversas instituições de dentro e fora do país, com palestras transmitidas ao vivo pelo canal Estatística UFF do YouTube (`www.youtube.com/estatisticauff`) e postadas no mesmo canal para posteriores visualizações. Três minicursos síncronos ocorreram durante o evento e os trabalhos aceitos para o evento foram transmitidos em formato de vídeo também via YouTube.

Como nas edições anteriores, o principal objetivo do evento foi o de criar um ambiente em que discentes e docentes, da UFF e de outras instituições, interagissem de forma a ampliar e complementar experiências acadêmicas e profissionais na área de Estatística. Consideramos que esse objetivo foi alcançado.

Departamento de Estatística

O Departamento de Estatística (GET), que está situado no Instituto de Matemática e Estatística da Universidade Federal Fluminense, é o responsável pela organização da 12ª Semana da Estatística, tendo como parceiros a Coordenação do Bacharelado em Estatística e o Laboratório de Estatística da UFF.

Comissão Organizadora

Estelina Capistrano	-	GET/UFF
Jessica Quintanilha Kubrusly	-	GET/UFF
Jony Arrais Pinto Junior	-	GET/UFF
Mariana Albi de Oliveira Souza	-	GET/UFF
Patrícia Lusié Velozo da Costa	-	GET/UFF
Rafael Santos Erbisti	-	GET/UFF

Programação Geral

PL: Palestra, MI: Minicurso.

Segunda-feira, 18 de Outubro

14:00–14:50	PL	**Alexandra Mello Schmidt** McGill University	Mapeamento bayesiano dos casos de Covid-19 ao longo dos bairros de Montreal, no Canadá
15:00–15:50	PL	**Leonardo Bastos** Fundação Oswaldo Cruz	Estatística aplicada a COVID-19 no Brasil
16:00–16:50	PL	**Joel M. C. da Rosa** Icahn School of Medicine at Mount Sinai	Aplicação da Estatística em Multi-Ômica
17:00–18:30	MI	**Ricardo Junqueira de Souza** Instituto de Segurança Pública do Rio de Janeiro	Uma jornada pelo tidyverse

Terça-feira, 19 de Outubro

14:00–14:50	PL	**Nathalia Demetrio** Banco Itaú e Insper	O que eu não aprendi em Estatística sobre Ciência de Dados
15:00–15:50	PL	**Mariana Raniere** Queen Mary University	Causal Dynamic Bayesian Networks for the Management of Glucose Control in Gestational Diabetes
16:00–16:50	PL	**Marcos Prates** UFMG	Estatística em Sociedade: Da metodologia a aplicações
17:00–18:30	MI	**Thiago Augusto Santos Lima** Globo	Sistemas de recomendação no R

Quarta-feira, 20 de Outubro

14:00–14:50	PL	**Franklin Torres** Superintendência de Pesq. e Estratégia de Segurança Pública do Ceará	A estatística na Segurança Pública do Ceará
15:00–15:50	PL	**Evandro Lopes** Cognitivo.ai	Cultura Analítica: Apresentando um case real de modelagem e desafios do mercado
16:00–16:50	PL	**Ana Oliveira** Dell Technologies	Inteligência Artificial, do que se alimenta, para onde vai e quais os dilemas éticos enfrenta?
17:00–18:30	MI	**Daniel dos Santos** Bacharelado em Estatística / UFF	Utilizando o R para Big Data, uma introdução prática ao SparkR

Segunda–feira, 18 de Outubro

Mapeamento bayesiano dos casos de Covid-19 ao longo dos bairros de Montreal, no Canadá

Alexandra Mello Schmidt PL

McGill University – Montreal, Canadá

Temos disponíveis o número de casos e o número de óbitos devido a covid-19 ao longo dos n=33 bairros da cidade de Montreal, no Canadá. Usualmente, este tipo de dado é modelado como seguindo uma distribuição de Poisson, cuja média é descrita pelo produto entre um offset e o risco relativo da doença. O log do risco é modelado como função linear de covariáveis, por exemplo, idade média do bairro, nível educacional médio do bairro, porcentagem de leitos para idosos, entre outras. No entanto, este modelo usual assume que média e variância são iguais, o que raramente é verdade para observações deste tipo. Usualmente, a variância é maior do que a média; este fenômeno chama-se sobredispersão. Discutirei diferentes formas de modelar o risco relativo de covid-19 ao longo dos bairros de Montreal. Começarei pelo modelo log-linear usual, mostrando os problemas com o uso deste modelo. Em seguida, apresentarei um modelo de mistura que inclui um efeito aleatório (não-observável) para capturar eventuais estruturas que estejam presentes nas observações e não são capturadas pelo modelo usual. Estes modelos de mistura lidam, naturalmente, com a questão da sobredispersão. As distribuições a priori deste efeito aleatório consideram desde independência entre bairros até uma estrutura espacial. O procedimento de inferência das quantidades desconhecidas dos modelos propostos segue o paradigma de Bayes. Também discutirei um modelo conjunto para o número de casos de covid-19 e o número de óbitos devido à covid-19 condicionado ao total de casos de cada bairro. Este trabalho foi desenvolvido como uma iniciação científica de Leo Vanciu sob minha orientação e de Victoire Michal, doutoranda em Bioestatística na Universidade McGill.

Apresentação disponível em:
www.youtube.com/watch?v=zNWTYJeRsjg&list=PLtjtKxC5uk9duYTICvDOTQwcQ1fEcFZKt&index=1&t=5s

Estatística aplicada a COVID-19 no Brasil

Leonardo Bastos PL

Fundação Oswaldo Cruz – Rio de Janeiro, Brasil

Nesta palestra vou apresentar resultados de projetos desenvolvidos pelo grupo de métodos analíticos em vigilância epidemiológica que vêm desde bem antes da epidemia monitorando os casos de hospitalizações pela síndrome respiratória aguda grave (SRAG) no sistema InfoGripe, como mostramos que o perfil de casos de SRAG mudou antes da Covid-19 ser formalmente monitorada, como calculamos os fatores de riscos para o agravamento da Covid-19, como seguimos o monitoramento da Covid-19 usando métodos de correção de atraso de notificação, e quais os próximos passos.

Apresentação disponível em:
www.youtube.com/watch?v=zHk4SDy9Jkw&list=PLtjtKxC5uk9duYTICvD0TQwcQ1fEcFZKt&index=2&t=193s

Aplicação da Estatística em Multi-Ômica

Joel M. C. da Rosa PL

Icahn School of Medicine at Mount Sinai – Nova York, Estados Unidos da América

Novas tecnologias têm permitido o aumento na compreensão de doenças e consequentemente a formulação de medicamentos mais eficazes e seguros. O caminho para a chamada Medicina de Precisão tem sido pavimentado com uma quantidade cada vez maior de dados biológicos coletados em diversos organismos. A multi-ômica, integração de dados moleculares como DNA, RNA, proteínas, metabólitos e microbioma, tem permitido um avanço sem precedentes na descrição de sistemas biológicos complexos. As Ciências da Computação, Biologia, Matemática e Estatística interagem nas diferentes etapas de análise desta extraordinária massa de dados. Apresentarei fundamentos de métodos estatísticos utilizados em dados provenientes da multi-ômica, especialmente a construção de variáveis latentes, a redução de dimensionalidade, algoritmos de predição e a visualização de dados. Incluirei aplicações em Dermatologia, em particular a utilização de dados multi-ômicos para compreender mecanismos relacionados à Dermatite Atópica e Psoríase.

Apresentação disponível em:
www.youtube.com/watch?v=56yqTqcRSMw&list=PLtjtKxC5uk9duYTICvD0TQwcQ1fEcFZKt&index=3&t=54s

Uma jornada pelo tidyverse

Ricardo Junqueira de Souza MI

Instituto de Segurança Pública – Rio de Janeiro, Brasil

O tidyverse é uma coleção de pacotes para ciências de dados construídos sob a filosofia tidy. Sob esta filosofia os pacotes devem obedecer 4 princípios: Reutilizar as estruturas de dados existentes, interligar funções simples com a utilização do pipe, abraçar a programação funcional e ser pensado para a utilização por seres humanos. Dentro deste conjunto de pacotes existem opções para importação de dados (readr), organização (dplyr, tidyr, tibble), manipulação de texto (stringr) data (lubridate) e por fim, visualização de dados (ggplot2). Neste minicurso vamos utilizar um conjunto de dados real para mostrar todo o caminho desde a importação dos dados, passando pelo tratamento e chegando a visualização dos mesmos usando apenas os pacotes do tidyverse.

Terça–feira, 19 de Outubro

O que eu não aprendi em Estatística sobre Ciência de Dados

Nathalia Demetrio PL

Banco Itaú e Insper - São Paulo, Brasil

Neste bate-papo iremos construir um entendimento sobre o que é a ciência de dados, considerados os principais momentos e necessidades que envolvem um fluxo de trabalho na ciência de dados, bem como a relação entre estes e o universo da Estatística.

Apresentação disponível em:
`www.youtube.com/watch?v=MHai21xrmEs&list=PLtjtKxC5uk9duYTICvDOTQwcQ1fEcFZKt&index=4&t=4s`

Causal Dynamic Bayesian Networks for the Management of Glucose Control in Gestational Diabetes

Mariana Raniere PL

Queen Mary University – Londres, Reino Unido

Patients suffering from chronic conditions may need to make frequent decisions about the management of their condition in partnership with their health professionals. However, this may not be possible as appointments are not always scheduled according to necessity but instead at a fixed frequency. Remote monitoring technology has the potential to generate patient data but without intelligent systems capable of analysing the data and offering advice, more data just increases the person's dependency on clinical staff for its interpretation. Decision-support systems that can give people more autonomy in the management of their condition can therefore benefit both the affected person and clinicians. We propose the use of Dynamic Bayesian Networks built from expert knowledge to interpret data and support decision-making, offering advice to patients suffering from a chronic condition. We argue that expert knowledge is needed as well as data to build such a decision-support system as the data that would be required to use machine learning will never be available in the current clinical system with all treatment decisions made at appointments scheduled at fixed intervals. We illustrate the methodology using a case study in Gestational Diabetes.

Apresentação disponível em:
www.youtube.com/watch?v=MHofgYSGhGU&list=PLtjtKxC5uk9duYTICvDOTQwcQ1fEcFZKt&index=5&t=10s

Estatística em Sociedade: Da metodologia a aplicações

Marcos Prates PL

UFMG – Belo Horizonte, Brasil

Nesse seminário irei fazer um passeio pelas diversas áreas que venho atuando. O principal objetivo é demonstrar que a pesquisa em estatística possibilita a solução de problemas reais desafiadores. Além disso, irei enfatizar a necessidade da interlocução da pesquisa com a indústria, ou seja, como é salutar a ponte entre esses meios na qual a transmissão do conhecimento gerado nas universidades é usado para resolver problemas práticos nas empresas. Com isso em mente, a grande coleta atual de dados e a necessidade de decisões assertivas baseadas em conhecimento, mostrar que os desafios reais geram e motivam a necessidade do desenvolvimento metodológico na nossa área.

Apresentação disponível em:
www.youtube.com/watch?v=QqIu5J-N68c&list=PLtjtKxC5uk9duYTICvDOTQwcQ1fEcFZKt&index=6&t=10s

Sistemas de recomendação no R

Thiago Augusto Santos Lima MI

Globo – Rio de Janeiro, Brasil

Sistemas de recomendação são técnicas que fornecem sugestões de itens a serem recomendados para um usuário. As sugestões fornecidas, nos sistemas de recomendação, visam ajudar os usuários em vários processos de tomada de decisão, bem como, quais itens comprar, quais músicas escutar ou quais notícias ler. A recomendação através da Filtragem Colaborativa é fundamentada na similaridade dos itens recomendados. A ideia básica é que, se um usuário gosta de um determinado item, também gostará de um item semelhante. Neste minicurso serão apresentados conceitos de um dos principais métodos de recomendação: a Filtragem Colaborativa. Além dos conceitos, será apresentado um novo pacote, o CFilt. Através dele poderão ser criados e construídos sistemas para realizar recomendações de filmes, séries, livros e muito mais. Também serão apresentados métodos para avaliar os desempenhos dos sistemas de recomendação e testar combinações que geram melhores resultados finais.

Quarta–feira, 20 de Outubro

A estatística na Segurança Pública do Ceará

Franklin Torres PL

Superintendência de Pesquisa e Estratégia de Segurança Pública do Ceará – Fortaleza, Brasil

Esta palestra pretende abordar a evolução da Estatística e do papel do estatístico no contexto da segurança pública do estado do Ceará, de sua implementação oficial às diretrizes atuais como órgão fundamental na SSPDS/CE.

Apresentação disponível em:
`www.youtube.com/watch?v=x7U-dzjcIEU&list=PLtjtKxC5uk9duYTICvDOTQwcQ1fEcFZKt&index=7`

Cultura Analítica: Apresentando um case real de modelagem e desafios do mercado

Evandro Lopes PL

Cognitivo.ai – São Paulo, Brasil

Atualmente é possível observar um movimento forte das empresas para se tornarem cada vez mais orientada a dados. Estimativas iniciais mostram que o mercado de dados vai adicionar US$13 trilhões na economia global até 2030. O objetivo desta apresentação é falar sobre um projeto de Ciência de Dados entregue pela Cognitivo.ai e compartilhar alguns desafios observados no mercado.

Apresentação disponível em:
www.youtube.com/watch?v=IKRgiJsl-xM&list=PLtjtKxC5uk9duYTICvD0TQwcQ1fEcFZKt&index=8

Inteligência Artificial, do que se alimenta, para onde vai e quais os dilemas éticos enfrenta?

Ana Oliveira PL

Dell Technologies – Rio de Janeiro, Brasil

Conversaremos sobre a importância dos dados para a construção de modelos de aprendizado de máquina, falaremos de algoritmos tradicionais, mas também de alguns novos que são particularmente interessantes quando não temos muitos dados, ou quando estes dados não podem ser compartilhados. Além disto discutiremos algumas questões éticas com respeito ao uso de IA e o que tem sido feito para endereçar algumas destas questões.

Apresentação disponível em:
www.youtube.com/watch?v=otvSaAJbbrI&list=PLtjtKxC5uk9duYTICvD0TQwcQ1fEcFZKt&index=9

Utilizando o R para Big Data, uma introdução prática ao SparkR

Daniel dos Santos MI

Bacharelado em Estatística / UFF – Niterói, Brasil

Seja ao assistir a um filme favorito, realizar uma compra online ou postar uma foto em uma rede social, são enviadas informação que poderão ser utilizadas posteriormente para recomendações e melhoria na experiência de usuário. Assim, tornou-se indispensável a capacidade de explorar grande volume de dados, das mais diversificadas fontes. Para resolver este desafio, foi desenvolvido o Apache Spark, um framework de código aberto que busca democratizar a utilização de técnicas de computação distribuída e a aplicação de diversas tarefas em grande escala. Com o Spark é possível ler, transformar, descrever e treinar modelos de aprendizado de máquina em bancos de dados onde técnicas tradicionais não são capazes de gerar resultados por limitação de recursos computacionais.

Neste minicurso o participante irá entender e aplicar algumas ferramentas encontradas no SparkR, a API do Apache Spark para a linguagem R, como intuito de resolver problemas comuns que são facilmente encontrados ao iniciar análises em big data.

Resumos Estendidos

Nesta seção apresentamos os resumos estendidos dos trabalhos aceitos na 12ª Semana da Estatística da UFF. Os trabalhos foram apresentados em forma de vídeo e estão disponíveis no canal Estatística UFF do YouTube (`www.youtube.com/estatisticauff`). O link para cada apresentação encontra-se no rodapé do respectivo resumo.

Influência de uma dieta contendo óleo de linhaça no metabolismo glicídico

Aline D'Avila Pereira (UFF)
Danielle Ribeiro (UFF)
Letícia Cardoso (UFF)
Carlos Alberto Soares da Costa (UFRB)
Gilson Teles Boaventura (UFF)
Luis Guillermo Coca Velarde (UFF)

Email de contato: alinedavila@id.uff.br, lgcocavelarde@id.uff.br, gilsontb@gmail.com.

Resumo

O óleo de linhaça é fonte de ácido alfa-linolênico que é importante para os estágios de crescimento e desenvolvimento corporal e atua no metabolismo glicídico. Sendo assim, o objetivo foi avaliar a influência do óleo de linhaça, em diferentes fases da vida, sobre a glicídico. Ao nascimento, houve a randomização em dois grupos: grupo controle (C) e grupo óleo de linhaça (OL). Aos 21 dias, os ratos foram desmamados e houve uma nova divisão: dos 18 animais do C, 9 continuaram recebendo C e os outros 9 passaram a receber OL; dos 18 animais do OL, 9 continuaram OL e os outros 9 passaram a receber C. Foi analisado a massa da ninhada; a massa e comprimento corporal; e ingestão alimentar. Quando os animais completaram 60 dias, foi realizado o teste oral de tolerância à glicose. Aos 67 dias, houve a eutanásia por punção cardíaca. No soro, foram analisados a glicemia de jejum, a insulina e o homeostasis model assessment - insulin resistance (HOMA-IR). O pâncreas foi coletado e analisado quanto à massa absoluta e reativa e quanto à análise histomorfométrica em relação à área da ilhota pancreática. Foi utilizado o teste ANOVA de dois fatores, sendo o nível de significância de 0,05. O consumo de OL durante a lactação promoveu ($p<0,05$): menor massa corporal. No período pós-lactação, observou-se ($p<0,05$): maior tolerância à glicose. Assim, o estudo demonstra que o óleo de linhaça no período pós-lactação pode atuar na prevenção de doenças crônicas.

Palavras-chave: Óleo de linhaça, Ratos, Glicemia, Insulina, Pâncreas.

Introdução

Programação metabólica é caracterizada por estímulos períodos críticos que afetam a longo prazo as respostas fisiológicas, metabólicas e genéticas do adulto [9]. Assim, alterações nesse período pode promover a prevenção de doenças crônicas não transmissíveis (DCNT), enfermidades que se desenvolvem ao longo da vida, consideradas um problema de saúde pública, visto alta prevalência de óbitos [17]. Contudo, segundo o Ministério da Saúde (MS), as DCNT podem ser prevenidas a partir de uma alimentação saudável durante toda a vida, inclusive, na infância e adolescência [2].

Essas fases são caracterizadas pelo estágio de crescimento e desenvolvimento, logo, é importante que a mãe, a criança e o adolescente tenham uma dieta rica em ácido linoleico (18:3 n-6, LA) e ácido alfa-linolênico (18:3 n-3, ALA), pois estes são necessários para o desenvolvimento dos tecidos [6]. Além disso, estudos experimentais e epidemiológicos avaliaram que AGPI atuam no metabolismo glicídico, visto que o n-3 atua no aumento da sensibilidade à insulina e, com isso, pode promover maior captação de glicose [4, 18].

Nesse contexto, o óleo de linhaça, extraído da parte interna da semente de linhaça, é considerado um alimento com propriedades funcionais, pois representa uma das maiores fontes de ALA do reino vegetal e baixa razão de n-6/n-3, apresentando 15% de LA e 56% de ALA [5, 16]. Sendo assim, a proposta desse estudo foi avaliar a influência de uma dieta contendo óleo de linhaça ofertada em diferentes fases da vida sobre metabolismo glicídico de ratos machos jovens.

Material e métodos

O presente estudo foi submetido e aprovado pela Comissão de Ética no Uso de Animais com o número: 892. Para isso, as rações utilizadas ao longo do experimento foram preparadas na Faculdade de Nutrição da UFF e seguiram as recomendações da AIN-93G [14].

No nascimento dos filhotes (P0), a massa corporal da ninhada foi mensurada. No mesmo período, as mães e suas ninhadas foram divididas em dois grupos: grupo controle (n = 18 filhotes machos) e grupo óleo de linhaça (n = 18 filhotes machos). Ambos os grupos foram acompanhados até completarem 21 dias, quando ocorreu uma nova divisão que caracterizou a dieta pós-lactação: dos 18 animais, cujas ratas lactantes foram alimentadas com dieta controle: 9 continuaram recebendo a dieta controle até completarem 67 dias; enquanto os outros 9, passaram a receber dieta contendo óleo de linhaça até completarem 67 dias; e dos 18 animais cujas ratas lactentes foram alimentadas com dieta contendo óleo de linhaça: 9 continuaram recebendo a dieta contendo óleo de linhaça até completarem 67 dias; enquanto os outros 9, receberam dieta controle até completarem 67 dias.

O consumo alimentar (g) durante todo o experimento foi aferido três vezes por semana. A massa (g) e comprimento (cm) dos filhotes foi aferida 3 vezes por semana, a partir do 21º até os 67º dias de idade. Quando os animais completaram 60 dias, foi realizado o teste oral de tolerância à glicose (TOTG) [3, 12]. Esse teste consiste em analisar a glicemia capilar após a administração oral de determinada quantidade de glicose [7]. No presente estudo, os animais ficaram 6 horas em jejum e, depois desse período, receberam, por gavagem, 1g de dextrose/kg de massa corporal e a glicemia periférica (caudal) foi observada no tempo zero (antes da administração), 15, 30, 60 e 120 minutos após a administração. Para a obtenção da solução que foi administrada, houve a dissolução de dextrose em soro fisiológico na proporção de 1:1. Esse dado foi analisado a partir da área sob a curva (AUC).

Quando os animais completaram 67 dias, houve a eutanásia e o sangue foi coletado por punção cardíaca até exsaguinação total. No soro, foi avaliado a glicemia de jejum (mg/dL) e a insulina (ng/dL), com esses dados foi calculado a resistência à insulina através do modelo matemátco: HOMA-IR (homeostasis model assessment - insulin resistance) = (insulina de jejum (UI/ml) x glicemia de jejum (mmol/l)) / 22,5 [10].

Após a coleta do sangue, o pâncreas foi retirado, limpo e pesado em balança analítica (precisão 0,0001 Bosch S2000, Brasil), para análise da massa absoluta (g) e relativa (g/100g). Uma amostra de pâncreas foi fixada em formol 10% e submetida aos processos de desidratação em uma série crescente de álcool etílico, xilol e inclusão em parafina para a avaliação da área da ilhota pancreática. A análise foi realizada no laboratório de técnica histológica, localizado no Departamento de Ciências Fisiológicas/UERJ. A avaliação foi determinada a partir da análise de imagens digitais adquiridas pela câmera digital Olympus DP72 acoplada ao microscópio Olympus BX51 com ocular 20X. A análise foi realizada com o auxílio de programa Image J – Pro-versão 4.5.0.29 (National Institute of Health, USA).

Os dados relacionados à ingestão alimentar, à massa, ao comprimento corporal, à tolerância à glicose, à glicemia de jejum, à insulina e ao HOMA-IR foram avaliados usando o método de análise de variância (ANOVA) de dois fatores (dieta e período de oferta da dieta: lactação ou pós-lactação). Para os dados referentes ao peso ao nascer foi utilizado o teste t-student. Foi considerado o nível de significância de 0,05 e a análise estatística foi realizada usando o software R versão 3.3.2.

Resultados e discussão

Não houve diferença significativa quando se comparou o consumo alimentar (dieta lactação: p=0,8935 e dieta pós-lactação: p=0,9896), a massa corporal da ninhada (p=0,9213), massa corporal no período pós-lactação (p=0,9708) e o comprimento corporal (dieta lactação: p=0,9285 e dieta pós-lactação: p=0,9448). Entretanto, a massa corporal foi menor a partir do 60º dia no grupo cujas mães consumiram óleo de linhaça durante a lactação (p=0,0001). Em relação ao consumo, o re-

sultado era esperado pois as rações eram isolipídicas, isoglicídicas, isoproteicas e isocalóricas, além de apresentarem a mesma quantidade e fonte de fibras. Além disso, vale ressaltar que as rações ofertadas apresentavam diferença somente no conteúdo lipídico, visto que o óleo de soja, utilizado na confecção da ração controle, é fonte de n-6 e o óleo de linhaça, da ração óleo de linhaça, é fonte de n-3.

O comprimento corporal não apresentou diferença significativa entre os grupos, no entanto, a massa corporal foi menor a partir do 60º dia até o final do experimento nos animais que consumiram óleo de linhaça no período de lactação. Isso pode ser explicado pelo fato de o óleo de linhaça ser fonte de ácidos graxos da família no n-3, que possuem efeito antiobesogênico [11].

O consumo de óleo de linhaça no período pós-lactação promoveu melhor tolerância à glicose (p=0,025), uma vez que os animais apresentaram menor área sob a curva. No entanto, não foi observada diferença significativa nas outras análises. Estudos têm elucidado que a diminuição da glicemia pós-prandial está associada a alimentos com fibras [8, 15, 13]. Neste estudo, as dietas apresentavam a mesma quantidade e fonte de fibras, sendo a diferenças em relação ao n-3 uma possível explicação para esse achado. O n-3 apresenta efeitos benéficos à função e sobrevida das ilhotas pancreáticas e atua diminuindo as concentrações de PG2, um composto relacionado com a maior secreção de insulina, assim, há aumento da sensibilidade à insulina e maior tolerância à glicose [4, 18, 1]. Assim, os achados do presente estudo demonstram efeitos significativos da dieta contendo óleo de linhaça nos estágios de lactação, infância e adolescência sobre o metabolismo glicídico.

Referências

[1] Baynes, H. W., Mideksa, S. e Ambachew, S. The role of polyunsaturated fatty acids (n-3 pufas) on the pancreatic b-cells and insulin action. *Adypocite 7* (2018).

[2] Brasil. Ministério da saúde. vigilância de doenças crônicas não transmissíveis (dcnt). 2018.

[3] Correia-Santos, A. M., Suzuki, A., Vicente, G. C., dos Anjos, J. S., Pereira, A. D., Lenzi-Almeida, K. C. e Boaventura, G. T. Effect of maternal use of flaxseed oil during pregnancy and lactation on glucose metabolism and pancreas histomorphometry of male offspring from diabetic rats. *Diabetes Res Clin Pract 106* (2014).

[4] Flachs, P., M.Rossmeisl e Kopecky, J. The effect of n-3 fatty acids on glucose homeostasis and insulin sensitivity. *Physiol Res 63* (2014), S93–S118.

[5] Goyal, A., Sharma, V., Upadhyay, N., Gill, S. e Sihag, M. Flax and flaxseed oil: an ancient medicine & modern functional food. *J Food Sci Technol 51* (2014).

[6] Green, K. H., Wong, S. C. F. e Weiler, H. A. The effect of dietary n-3 long-chain polyunsaturated fatty acids on femur mineral density and biomarkers of bone metabolism in healthy diabetic and dietary-restricted growing rats. *Prostaglandins, Leukot Essent Fatty Acids 71* (2004), 121–130.

[7] Gross, J. L. Diabetes melito: Diagnóstico, classificação e avaliação do controle glicêmico. *Arq Bras Endocrinol Metabol 46* (2002).

[8] Kapoor, S., Sachdeva, R. e Kochhar, A. Efficacy of flaxseed supplementation on nutrient intake and other lifestyle pattern in menopausal diabetic females. *Ethnomediciner 5* (2011).

[9] Langley-Evans, S. C. Nutrition in early life and the programming of adult disease: a review. *J Hum Nutr Diet 28* (2015), 1–14.

[10] Matthews, D. R., Hosker, J. P., A. S. Rudenski, B. A. N., Treacher, D. F. e Turner, R. C. Homeostasis model assessment: insulin resistance and b-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia 28* (1985).

[11] Mohammadi-Sartang, M., Ghorbani, M. e Mazloom, Z. Effects of melatonin supplementation on blood lipid concentrations: A systematic review and meta-analysis of randomized controlled trials. *Clin Nutr 37* (2018).

[12] Olvera-Hernández, V., Ble-Castilho, J. L., Betancur-Ancona, D., Acevedo-Fernández, J. J., Castellanos-Ruelas, A. e Chel-Guerrero, L. Effects of modified banana (musa cavendish) starch on glycemic control and blood pressure in rats with high sucrose diet. *Nutr Hosp 35* (2018).

[13] Prasad, K. e Dhar, A. Flaxseed and diabetes. *Curr Pharm Des 22* (2016).

[14] Reeves, P. G., Nielsen, F. H. e Jr, G. C. F. Ain-93 purified diet of laboratory rodents: final report of the american institute of nutrition ad hoc writing committee on the reformulation of the ain-76a rodents diet. *J Nutr 123* (1993).

[15] Rhee, Y. e Brunt, A. Flaxseed supplementation improved insulin resistance in obese glucose intolerant people: a randomized crossover design. *Nutr J 10* (2011).

[16] Richardson, C. E., Hennebelle, M., Otoki, Y., Zamora, D., Yang, J., Hammock, B. D. e Taha, A. Y. Lipidomic analysis of oxidized fatty acids in plant and algae oils. *J Agric Food Chem 65* (2017).

[17] WHO. World health organization. global status report on non communicable diseases 2010. 2010.

[18] X, X. W. e Chan, C. B. n-3 polyunsaturated fatty acids and insulin secretion. *J Endocrinol 224* (2015).

O uso da Correlação de Postos de *Spearman* como Determinação da quantidade de grupos para Análise de *Cluster*

Carla Cristina Passos Cruz (UERJ)
Regina Serrão Lanzillotti (UERJ)

E-mail de contato: carlapassos2889@gmail.com, reginalanzillotti@gmail.com.

Resumo

O tratamento dos textos corresponde a um conjunto de ações relacionadas aos documentos para a extração cognitiva. O fato de uma coletânea textual possuir muitas palavras faz com que as mesmas não possuam frequências iguais e, diante este fato utilizou-se medidas para verificação do grau de relacionamento das palavras. Os três documentos escolhidos abordaram o tema nutrigenômica e a opção metodológica para definir a quantidade de grupos para o Agrupamento utilizou-se a correlação ordinal de *Spearman*. Analisando os resultados, os Documentos 2 e 3 apresentaram uma associação mais relevante embora não distanciando da correlação dos Documentos 1 e 3. O gráfico tridimensional indicou que o termo "alimento" distanciava-se dos outros. O método proposto atingiu o objetivo de identificar cenários segundo agrupamento de termos considerados correlacionados.

Palavras-chave: *Text Mining*, Agrupamento, Correlação de *Spearman*.

Introdução

A informação textual tem como representação termos linguísticos que podem trazer incerteza, envolvidos na resolução de um problema, que podem ser decorrentes de alguma informação faltante ou que tem mais de uma solução [5]. Por isso, há a necessidade da aplicação de técnicas que possibilitem a extração e análise de dados para propiciar opiniões que abranjam temáticas textuais. Dentre as abordagens utilizadas em Mineração de Texto, se destaca a Análise de *Cluster*, utilizado pela Recuperação da Informação (RI), Aprendizado de Máquina (AM), Inteligência Computacional (IC), dentre outros.

O método é bastante utilizado para a identificação de conteúdos similares, caso não se tenha definição dos assuntos tratados em cada texto e se deseja separá-los por assunto [11]. Há duas formas possíveis de classificá-lo que são os Métodos Hierárquicos, que são técnicas simples onde os dados são divididos sucessivamente e não requerem uma definição da quantidade de grupos; e Métodos Não-Hierárquicos, que são utilizados com o objetivo de encontrar diretamente uma partição dos N elementos em g grupos, onde cada partição é um *Cluster*.

No entanto, para a escolha da quantidade de grupos, não há um consenso quanto a uma aplicação que a faça, ficando a cargo do pesquisador escolher. Diante deste fato, o presente trabalho tem como objetivo propor a determinação da quantidade de grupos pela Correlação Ordinal (ou de Postos) de *Spearman*, que não exige a suposição de que a relação entre as variáveis seja linear, nem requer que as mesmas sejam quantitativas [1].

Materiais e Métodos

A Mineração de Texto exige o *KDT - Knowledge Discovery in Text*, que consiste em um conjunto de procedimentos para extrair e recuperar dados textuais considerados relevantes, composto pelas etapas, ilustradas na Figura 1 [3] [8]. Primeiramente realiza-se a coleta e armazenamento dos dados que são analisados, que são conhecidos na literatura como *corpus* ou *corpora*. Em seguida, os textos passam pelo pré-processamento, considerada a etapa mais importante do processo, pois efetua a limpeza, padronização e seleção dos termos, além da criação da matriz termo-documento de linhas e colunas.

Figura 1: Etapas do *KDT*.
Fonte: As autoras, 2021.

Mas, antes da aplicação do algoritmo de agrupamento não hierárquico, faz-se necessária a definição da quantidade de grupos. Gath e Geva (1989) [4] descreveram três requisitos que servem como critério para se definir uma partição, isto é, uma quantidade de grupos aceitável, Figura 2.

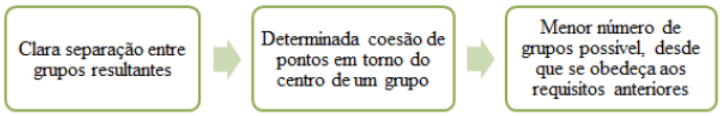


Figura 2: Requisitos para Partição.
Fonte: As autoras, 2021.

Neste trabalho a determinação da quantidade de grupos foi obtida pela Correlação de Postos de *Spearman* r_s, de acordo com a conformidade entre as informações de particionamento e distância [6]. Primeiro são obtidas as frequências absolutas $f(x)_{abs}$ pela matriz termo-documento e, em seguida, obtêm-se as frequências relativas $f(x)_{rel}$ em cada documento e a conjunta. Depois, ordenam-se as frequências relativas $f(x)_{rel}$ obtidas em cada documento através do ranqueamento onde são obtidos os chamados postos. Uma vez ranqueadas, as mesmas indicam as coordenadas do espaço tridimensional dos postos advindos do ranque do termo em cada documento a ser usado no método não-hierárquico, cuja expressão toma fórmula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

onde:

- d, diferença entre os postos (frequências), que são calculados par a par (v_i, y_i), (v_i, w_i) e (w_i, y_i), ou seja, os postos v_i dentre os valores de v e assim por diante;
- n, número de elementos (termos).

Esta expressão será aplicada a cada par de documentos gerando três correlações: Documento 1 e Documento 2; Documento 1 e Documento 3; Documento 2 e Documento 3, possibilitando a construção do gráfico de dispersão.

Resultados e Discussão

Foram selecionados três textos da área de Nutrigenômica (sites Sophie Deram [2], NutMed [7] e Profissão BioTec [9]), que é uma área que torna possível o estudo das interações entre dieta, nutrientes e genes e essa área do conhecimento pode ser considerada como ferramenta para terapia

nutricional, pois são essenciais para entender como os nutrientes modulam *in vivo* os mecanismos das doenças crônicas [10]. Para a aplicação da medida de associação proposta, onde se obteve um total de 551 termos. Em seguida, teve o pré-processamento que se iniciou pela Indexação desenvolvida mediante o *software RStudio* aplicada a cada documento, que converteu letras maiúsculas em minúsculas e retirou acentos, caracteres especiais, números, espaços extras, *links*, figuras e emojis, *stopwords* e a Normalização (união dos termos com o mesmo radical e termos sinônimos).

Os respectivos termos, oriundos dos textos escolhidos, podem ser considerados como covariáveis, pois permitem a construção de uma tabela de contingência, em que as células correspondem as frequências observadas dos termos em cada documento. Na etapa seguinte foi construído o histograma dos percentuais referente aos três Documentos que permitiu delimitar o corte limiar indicando em $0,36\%$ totalizando 90 termos, Figura 3.

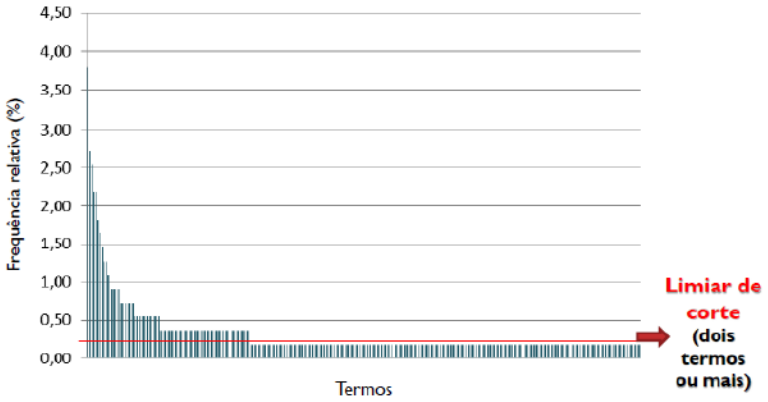


Figura 3: Seleção dos Termos.
Fonte: As autoras, 2021.

O próximo passo corresponde obter a correlação ordinal de *Spearman*, Tabela 1, uma vez que a frequência relativa foi traduzida em escala ordinal em escores crescentes para os termos em cada documento. As correlações ordinais permitiram criar uma matriz onde se identificou os valores $0,1890$, $0,0362$ e $-0,0585$, para os pares de Documentos 1 e 2, Documentos 1 e 3 e Documentos 2 e 3, respectivamente.

Apesar das correlações não terem apresentado valores expressivos, observa-se que o conjunto de termos nos Documentos 1 e 2 apresentaram uma associação mais relevante, os Documentos 1 e 3 apresentaram uma correlação mínima, e que a correlação entre os Documentos 2 e 3 se mostrou negativa, ou seja, correlação inversa.

Tabela 1: Correlação de *Spearman* entre os documentos

	Documento 1	**Documento 2**	**Documento 3**
Documento 1	$1,0000$	$0,1890$	$0,0362$
Documento 2	$0,1890$	$1,0000$	$-0,0585$
Documento 3	$0,0362$	$-0,0585$	$1,0000$

Fonte: As autoras, 2021.

O gráfico de dispersão tridimensional correspondentes aos três documentos e 90 termos selecionados pelo limiar de corte referente a frequência relativa, viabilizaram obter uma orientação para estabelecer o número de grupos. O gráfico tridimensional, Figura 4, indicou a discriminação do termo "alimento" e "nutrigenoma", ambos distanciado dos demais, embora "alimento" teve o maior destaque. Os outros quatro grupos indicaram que três deles mostraram-se com pouquíssimo distanciamento: "genoma"; "gene " e "expressao"; "estudo", "dieta" e "nutricao".

O último agrupamento apresentou superposição de termos o que mostrou a ausência de discriminação, um vez que agregou termos como "brasil", "individuo", "nutriente", "ciencia", "populacao", "vida", "remedio", "ligacao", "europeu", "presente", etc., não constituindo um cenário explícito relacional. O método proposto atingiu o objetivo de identificar cenários segundo agrupamento de termos considerados correlacionados. Esta opção de definição de quantidade de grupos serviu como insumo para aplicação dos métodos não-hierárquicos *Fuzzy C-Means* e *Fuzzy C-Medoids*.

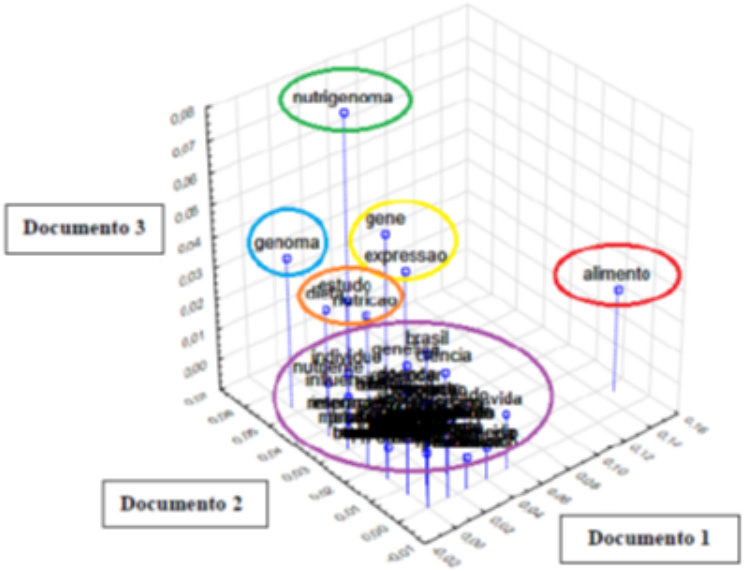


Figura 4: Gráfico de Dispersão Tridimensional das frequências relativas no sentido de estabelecer o número de agrupamentos.

Fonte: As autoras, 2021.

Agradecimentos

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

Referências

[1] Aranha, C. N. *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional.* Tese (Doutorado em Engenharia Elétrica), Pontifícia Universidade Católica do Rio de Janeiro, 2007.

[2] Deram, S. O que é nutrigenômica? *Sophie Deram – Doutora da USP e coaching em Nutrição* (2016). Disponível em: `https://www.sophiederam.com/br/nutricoaching/nutrigenomica/`. Acesso em: 20 set. 2021.

[3] Dixon, M. An overview of document mining technology. *Unpublisher paper* (1997).

[4] Gath, I. e Geva, A. B. Unsupervised optimal fuzzy clustering. *IEEE Transactions Pattern Analysis and Machine Intelligence* (1989).

[5] Goularte, F. B. *Método fuzzy para a sumarização automática de texto com base em um modelo extrativo (FSumm).* Dissertação (Mestrado em Ciência da Computação), Universidade Federal de Santa Catarina, 2015.

[6] Handl, J., Knowles, J. e Kell, D. B. Computational cluster validation in post-genomic data analysis. *Journal Bioinformatics, Maryland* (2005).

[7] NUTMED, E. O que é a nutrigenômica. *Site NutMed* (s.d.). Disponível em: `https://nutmed.com.br/blog/novidades/noticia-67`. Acesso em: 20 set. 2021.

[8] Tan, A.-H. Text mining: The state of the art and the challenges. *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases* (1999).

[9] Tonial, G. Nutrigenômica: nutrição a nível molecular. *Site Profissão Biotec* (2016). Disponível em: `https://profissaobiotec.com.br/nutrigenomica-nutricao-a-nivel-molecular/`. Acesso em: 20 set. 2021.

[10] Valente, M. A. S., de Albuquerque Barbosa, M. C., Rodrigues, C. V., Vieira, P. A. F. e de Oliveira Barbosa, M. Nutrigenômica/nutrigenética na elucidação das doenças crônicas.

[11] Wives, L. K. *Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva*. Tese (Doutorado em Computação), Universidade Federal do Rio Grande do Sul, 2015.

Técnicas de Mineração de Texto e de Análise de Conglomerados aplicadas em banco de dados de automóveis

Danielle Ribeiro Pereira da Silva (UFF)
Jessica Quintanilha Kubrusly (UFF)

Email de contato: daniellerps@id.uff.br, jessicakubrusly@id.uff.br.

Resumo

O banco de dados de dados deste trabalho é um banco de textos com informações sobre modelos de automóveis novos extraídos da *Internet* via *web scraping*. O objetivo é aplicar técnicas de Mineração de Texto e Análise de Conglomerados a fim de agrupar documentos referentes a automóveis com as mesmas características e assim poder criar indicadores de preços de forma automática. A validação dos resultados foi feita pelas Nuvens de Palavras.

Palavras-chave: Mineração de Texto, Análise de Conglomerados, Dados não-supervisionada, Nuvem de Palavras, Automóveis Novos.

Introdução

O avanço da tecnologia e do acesso à *Internet* facilitou e modernizou o modo do consumidor realizar compras. Isso é consequência das inúmeras ofertas que o consumidor é exposto pelo acesso à *Internet*, através de *sites* ou aplicativos. Diante dessa mudança de perfil dos consumidores, tornou-se interessante criar indicadores que pudessem medir o consumo via *Internet*. Com o intuito de melhor representar o perfil dos consumidores, o IBRE (Instituto Brasileiro de Economia), órgão responsável pela criação de indicadores econômicos na Fundação Getúlio Vargas, modificou a coleta de preços, antes feita apenas presencialmente, agora incluindo também preços coletados via *Internet*. Para isso, são utilizadas técnicas de *web scraping*, que consistem em automatizar a extração de dados da *Internet*, que coletam mais de 5 milhões de preços mensalmente.

Um enorme desafio no processamento dos dados coletados via *Internet* dá-se no agrupamento de itens semelhantes, que acabam sendo diferenciados pela forma que cada *site* os apresenta, seja pela ordem da escrita presente na sua descrição, ou por outros motivos. Diante desse cenário, sentiu-se a necessidade de realizar algum tratamento nos dados textuais a fim de aproveitá-los para o cálculo do indicador em questão.

Visando solucionar esse problema, a proposta deste estudo é processar, analisar e classificar em diferentes grupos, via técnicas de Mineração de Texto e de Análise de Conglomerados, um banco textual. Trata-se de um banco de dados semi-estruturado, composto por características de automóvel novos ofertado em diferentes *sites*. Tais características aparecem em formato de texto livre e também como variáveis que complementam essa informação, como o ano de fabricação, motor, cor, tipo de combustível, entre outras, que diferenciam o preço do automóvel. A informação contida nessas variáveis complementares são diferentes para diferentes *sites* e ainda contém grande parte delas não preenchidas, por isso o campo textual apresenta a informação de maior interesse.

Material e métodos

O material utilizado neste trabalho é um banco de documentos. Para o trabalho em questão, os documentos contém textos com a descrição de modelos de automóveis novos ofertados em diferentes *sites* de compra e venda na *Internet*. Esses documentos serão divididos em dois grupos: banco de treino e banco de teste. Vamos considerar que o banco de treino, aquele usado para ajustar os métodos descritos, contém n documentos. Já o banco de teste será utilizado somente para a validação dos resultados.

Mineração de Texto

A Mineração de Texto é uma área particular da Mineração de Dados que tem como finalidade encontrar padrões e extrair informações presentes em um banco textual. É um processo que utiliza diversos algoritmos de aplicação e pode ser dividido em três etapas principais: pré-processamento dos dados; transformação do banco de dados textual em banco de dados numérico; e a extração de informações importantes através de métodos de aprendizados numéricos.

A etapa de pré-processamento dos dados tem como objetivo obter alguma estrutura para o banco textual. Trata-se de um processo não trivial, uma vez que pode ser necessário aplicar técnicas combinadas até chegar ao resultado desejado [4]. Dentre as diversas técnicas existentes, neste trabalho serão abordadas algumas delas, descritas brevemente a seguir.

Tokenização: A *tokeninzação* é a primeira etapa do pré-processamento e seu objetivo é extrair unidades mínimas do texto a partir de um texto livre. Essas unidades são chamadas de *tokens*.

Remoção de *stop words*: *stop words* são palavras de ocorrência frequente em um idioma que não agregam informação relevante a um texto. Exemplos de *stop words* são pontuação, artigos, preposições e conjunções, palavras com pouca informação a ser agregada ao texto num geral. Além disso, às vezes, pronomes também são considerados *stop words*. Essa etapa do pré-processamento é importante, pois, retira do banco textual a maior parte dos dados que não são importantes para a extração de informação.

Normalização: Segundo Carrilho [4], a Normalização é a técnica de redução de léxico que se baseia no agrupamento de *tokens* que compartilham de um mesmo padrão. O objetivo nesta etapa é retirar o máximo de variações de uma mesma palavra existentes no banco textual. Após essa unificação os *tokens* passam a ser ser chamados de termos.

Seleção dos Termos: Esta etapa tem como objetivo reduzir o número de termos encontrados ao final da etapa de pré-processamento, obtendo assim um subconjunto mais preciso e representativo de termos. Um critério comum é selecionar os termos de acordo com a sua frequência. Baseando-se no método de Luhn [6], os termos de maior e menor frequência são julgados não relevantes e por isso o critério de seleção considera apenas os termos com frequência intermediária. Para esse trabalho serão desconsiderados os termos com frequência absoluta abaixo de 5% e os termos com frequência absoluta acima de 85% da quantidade total de documentos.

Matriz Termo-Documento: Após toda limpeza, unificação e seleção dos termos relevantes, a última etapa do pré-processamento consiste na construção da matriz termo-documento. Considerando o banco de treino com n documentos e que ao final do pré-processamento restaram m termos, a matriz termo-documento será uma matriz de dimensão $n \times m$ onde cada elemento a_{ij} terá valor 1 se o termo j aparece no documento i e zero caso contrário.

Análise de Conglomerados

A Análise de Conglomerados é um amplo conjunto de técnicas de classificação não supervisionada que consiste em construir agrupamento das unidades amostrais do banco de dados de acordo com algum critério de semelhança entre as variáveis explicativas. O objetivo é construir grupos com propriedades homogêneas a partir de grandes amostras heterogêneas [2]. Sendo assim, esses grupos não devem se sobrepor, cada elemento deve pertencer a um e apenas um grupo, e dentro do mesmo grupo, os elementos devem ser relativamente próximos uns dos outros, certamente muito mais próximos do que os elementos de outros grupos [3].

Para a aplicação de uma análise de conglomerados suponha n observações de m variáveis explicativas. Seja x_{ij} a i-ésima observação da j-ésima variável. Vamos chamar de $\mathbf{X}_i = (X_{i1}, \ldots, X_{im}) \in \mathbb{R}^m$ o vetor com todas as i-ésimas observações das variáveis X_j, $j = 1, 2, \ldots, m$. Veja que temos n vetores desses, um para cada observação, e cada um desses vetores pode ser considerado um ponto no $\mathbb{R}^m$. O objetivo do método é definir conjuntos de vetores próximos (conglomerados) no $\mathbb{R}^m$.

Para o problema tratado neste trabalho temos: n é o número de documentos analisados; m o número de termos ao final do pré-processamento da Mineração de Texto; x_{ij} é a variável indicadora que assume valor 1 quando o termo j aparece no documento i e zero caso contrário. Veja que os valores de x_{ij} são encontrados na matriz termo documento. Além disso, o vetor $\mathbf{X}_i = (X_{i1}, \ldots, X_{im}) \in \mathbb{R}^m$, que é a i-ésima linha da matriz termo-documento, guarda o resultado da variável indicadora para cada termo j com relação ao documento i.

Neste trabalho será utilizada a técnica hierárquica aglomerativa [2]. Como as variáveis explicativas são variáveis indicadoras, será adotada a medida de similaridade entre elementos amostrais

conhecida como Coeficiente de Jaccard [2], apresentada na Equação 1 (a). Já a similaridade entre conglomerados será calculada pela ligação média [3], Equação 1 (b).

$$(a)\ \ s_{i,k} = \frac{a}{a+b+c} \qquad\qquad (b)\ \ s_{AB}^{C} = \frac{\sum_{i \in C_A} \sum_{k \in C_B} s_{i,k}}{N_{C_A} \times N_{C_B}} \tag{1}$$

sendo $s_{i,k}$ a similaridade entre $\mathbf{X_i}$ e $\mathbf{X_k}$, a o total de vezes em que $x_{ij} = x_{kj} = 1$, b o total de vezes em que $x_{ij} = 1$ e $x_{kj} = 0$, c é o total de vezes em que $x_{ij} = 0$ e $x_{kj} = 1$, s_{AB}^{C} a similaridade entre os conglomerados C_A e C_B, N_{C_A} e N_{C_B} o número de elementos amostrais nos conglomerados C_A e C_B, respectivamente.

Nuvem de Palavras

Para validar se os métodos aplicados neste trabalho geraram bons resultados, será utilizada a Nuvem de Palavras (ou *Wordcloud* em inglês). A Nuvem de Palavras é uma representação gráfica das palavras presentes em um conjunto de textos. O método parte do princípio da frequência de cada palavra presente no texto e que são apresentadas de diferentes tamanhos na imagem. O tamanho de cada palavra na Nuvem de Palavras corresponde à sua frequência, ou seja, palavras em maior evidência são aquelas que aparecem mais vezes no texto já as palavras menores tratam-se das palavras com pouca frequência.

Resultados e discussão

O banco de dados utilizado neste trabalho é composto pela descrição de modelos de automóveis novos ofertados em diferentes *sites* de compra e venda, como a descrição textual completa do modelo do automóvel, ano de fabricação, motor, tipo de combustível, cor, entre outras especificações de um automóvel. A coleta dessas informações foi feita de forma automatizada através de técnicas de *web scraping*, coletadas diariamente durante todo o ano de 2020.

Com o objetivo de reduzir o banco de dados, foi extraída apenas uma observação (ou modelo de automóvel) de cada opção existente no banco. O banco de dados foi dividido em dois pedaços. O primeiro deles, o banco de treino, com 80% das observações do banco completo, e o segundo pedaço, o banco de teste, com 20% das observações. Essa partição foi feita com o pacote *caret* [5] e realizada de forma aleatória.

Através da descrição textual do modelo do automóvel foi extraída e criada a variável correspondente a marca do automóvel, o que possibilitou a filtragem do banco de dados completo pela marca do automóvel. Nesse trabalho serão apresentados os resultados para a marca Peugout, mas tudo pode ser replicado para qualquer uma das marcas. Essa parte de manipulação do banco de dados foi realizada utilizando o pacote *dplyr* [10], que possui funções apropriadas para isso.

Como resultado, o banco de treino utilizado neste trabalho, formado por documentos com descrição de modelos de automóveis novos para a marca Peugout, contém 140 documentos, enquanto o banco de teste contém 20 documentos. A Figura 1(b) apresenta a nuvem de palavras considerando todos os documentos do banco de treino. Nela podemos ver termos que se referem a modelos distintos, como por exemplo: automático e manual; cores e anos específicos. Espera-se que tais termos sejam separados em grupos distintos.

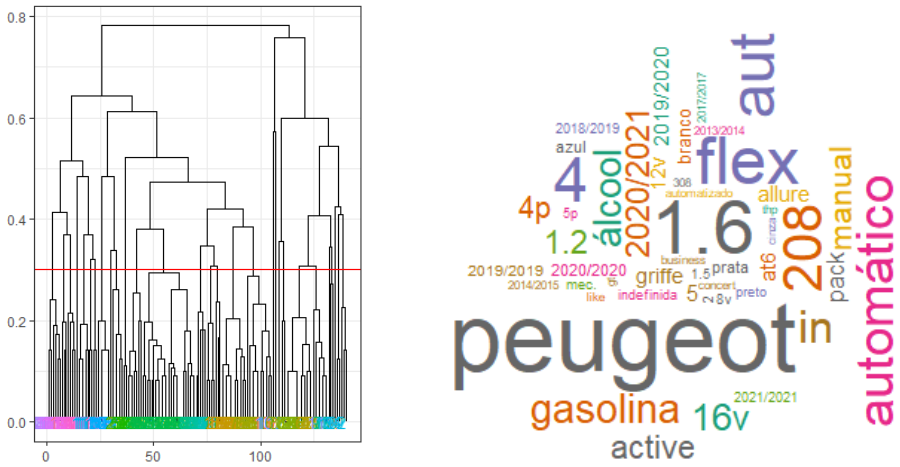
Na etapa de pré-processamento foi utilizado o Programa R [7] e o pacote *quanteda* [1], que possui funções que auxiliam no processo de mineração de textos. Logo na *tokenização* foi visto que existiam diversos *tokens* que representavam a mesma palavra, mas estavam escritos de forma abreviada ou com palavras similares. Devido a isso, foi necessário realizar uma etapa manual que passa por todos os *tokens* unificando aqueles que antes eram apresentados em diferentes formas.

Foram removidos dos documentos toda a acentuação, toda a lista de palavras fornecida pelo pacote *quanteda* [1], que são consideradas *stopwords* além de palavras encontradas no banco textual em questão que não apresentavam informação relevante como: "+","varias","outra" e "indefinida". Após a *tokenização* e a remoção dos *stopwords* os tokens resultantes desse processo apresentaram características únicas contendo também muitas palavras em inglês e com isso não foi necessário realizar a etapa de normalização.

A realização das etapas descritas gerou um total de 47 termos. Foram então removidos os termos de frequência acima de 85% e abaixo de 5%. Após a seleção dos termos restaram 22 termos. A matriz termo-documento foi criada com 140 linhas e 22 colunas.

Para a realização da Análise de Conglomerados e criação do dendograma foram utilizados os pacotes *stats* [7] e *ggdendro* [9]. A matriz de similaridade foi feita manualmente através de um laço (*loop*), que percorre cada posição preenchendo-a com a proximidade entre os documentos calculada através do Coeficiente de Jaccard, Equação 1 (a). Criada a matriz de similaridade, o passo seguinte foi a aplicação do algoritmo hierárquico aglomerativo e em seguida a aplicação do dendograma para a decisão do melhor número de grupos.

Através da inspeção visual do dendograma, Figura 1(a), foi escolhido ponto de corte na altura $0,3$, que gerou 29 conglomerados. As nuvens de palavras de cada um desses conglomerados encontra-se na Figura 2. Nessa figura são apresentadas duas nuvens por grupo, a da esquerda se refere à nuvem criada pelos documentos do banco de treino que pertencem ao conglomerado.



(a) Dendograma (b) Nuvem de Palavras

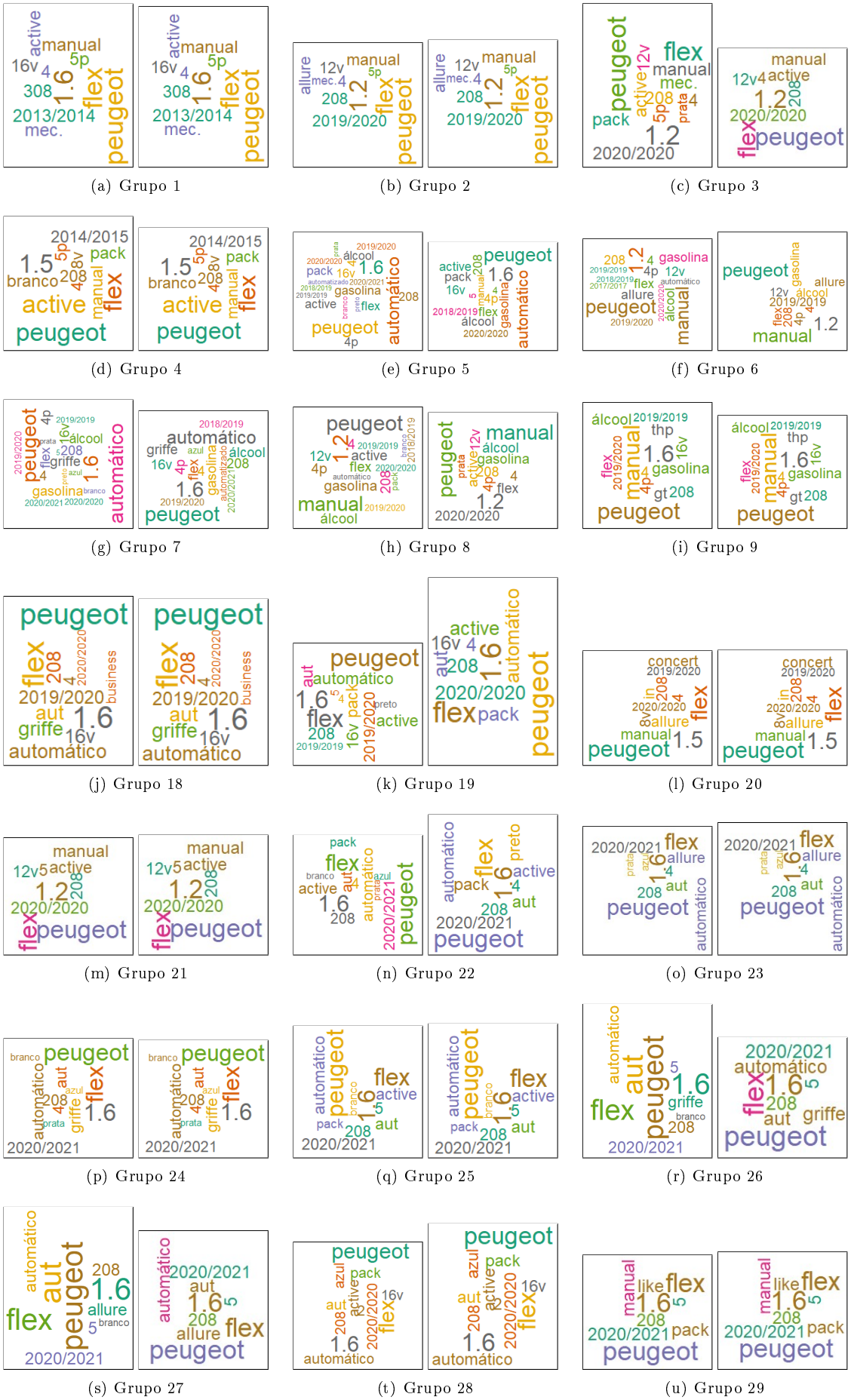
Figura 1: Resultados para o Banco de Treino

Uma vez definido os conglomerados, a partir do banco de treino, os documentos do banco de teste foram alocados em um dos 29 conglomerados da seguinte forma. Primeiro, para cada documento do banco de teste, é definido um vetor de variáveis indicadoras para a ocorrência ou não dos 22 termos da matriz termo-documento, criada na etapa de pré-processamento. Depois é encontrado o documento no banco de treino, representado por uma linha da matriz termo-documento, mais similar ao documento do banco de teste, representado pelo vetor de variáveis indicadoras. Para isso foi usado a função `knn` do pacote *class* [8]. O documento do banco de teste será alocado no conglomerado ao qual pertence este documento mais próximo a ele.

A fim de analisar se os conglomerados foram capazes de distinguir os modelos dos automóveis, tanto no banco de treino quanto no banco de teste, a Figura 2 apresenta as nuvens de palavras de alguns grupos: dos grupos 1-9 e dos grupos 18-29. Por causa do limite de espaço não foi possível apresentar as nuvens de palavras de todos os 29 grupos. São 2 nuvens por grupo, uma com os documentos do banco de treino (esquerda) e a outra com a adição dos documentos do banco de treino aos do banco de teste, que entraram no conglomerado em questão (direita).

Para analisar os resultados da Figura 2, veja primeiro as nuvens da esquerda, com as palavras dos documentos dos dados de treino. Como já comentado, palavras como automático e manual, versões do modelo (Like, Active, Allure ou Griffe), cores ou anos diferentes não devem aparecer no mesmo conglomerado, pois são termos conflitantes, referentes à modelos distintos. Percebemos que nos grupos 5-8, 10, 11, 15 aparecem dois ou mais pares de termos conflitantes. Já os grupos 9, 12, 13, 16, 17, 19, 20, 22-24 apresentam apenas um par de termos conflitantes. Os demais grupos (1-4, 14, 18, 21, 25-29), total de 12 grupos, não contém termos conflitantes. Esses últimos caracterizam bem um modelo de carro e os dados contidos em cada um desses conglomerados poderiam ser usados para a criação de um indicador de preço do modelo em questão.

Pensando agora no método como um classificador, a análise será concentrada nos 12 grupos que caracterizaram bem um modelo de automóveis. Em alguns deles as nuvens da direita são idênticas as da esquerda (grupos 1, 2, 4, 18, 21, 25, 28 e 29). Isso sugere que nenhum documento do banco de teste foi incluído nesses conglomerados. As outras nuvens, que tiveram alguma alteração, não deixaram de caracterizar o modelo do automóvel, indicando que a classificação ocorrida no banco de teste foi feita de forma correta nesses conglomerados.



(a) Grupo 1 (b) Grupo 2 (c) Grupo 3

(d) Grupo 4 (e) Grupo 5 (f) Grupo 6

(g) Grupo 7 (h) Grupo 8 (i) Grupo 9

(j) Grupo 18 (k) Grupo 19 (l) Grupo 20

(m) Grupo 21 (n) Grupo 22 (o) Grupo 23

(p) Grupo 24 (q) Grupo 25 (r) Grupo 26

(s) Grupo 27 (t) Grupo 28 (u) Grupo 29

Figura 2: Análise dos resultados através das nuvens de palavras.

Referências

[1] Benoit, K., K.Watanabe, H.Wang, P.Nulty, A.Obeng, S.Müller e A.Matsuo. quanteda: An r package for the quantitative analysis of textual data. *J. Open Source Softw. 3*, 30 (2018), 774.

[2] Härdle, W. K. e Simar, L. *Applied multivariate statistical analysis*. Springer, 2019.

[3] James, G., Witten, D., Hastie, T. e Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. 2017.

[4] Junior, J. R. C. Desenvolvimento de uma metodologia para mineração de textos. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro, 2008.

[5] Kuhn, M. *caret: Classification and Regression Training*, 2021.

[6] Luhn, H. P. The automatic creation of literature abstracts.

[7] Team, R. C. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[8] Venables, W. N. e Ripley, B. D. *Modern Applied Statistics with S*. Springer, 2002.

[9] Vries, A. e Ripley, B. *ggdendro:Create Dendrograms and Tree Diagrams Using 'ggplot2'*, 2020.

[10] Wickham, H., R.François, L.Henry e K.Müller. *dplyr:A Grammar of Data Manipulation*, 2021.

Modelos para dados de área com coeficientes variando espacialmente

Dayana Gimenes da Silva Ribeiro (UFF)
Ricardo Junqueira de Souza (UFF)
Jony Arrais Pinto Junior (UFF)

Email de contato: dayanagimenes@id.uff.br.

Resumo

Na modelagem de dados de área, usualmente, recorre-se a inclusão de efeitos aleatórios espacialmente estruturados, por meio de suas distribuições *a priori*. Em alguns cenários, somente a presença dos efeitos espaciais não consegue lidar com toda a heterogeneidade espacial existente no problema ou ainda é possível que o pesquisador deseje investigar como o efeito de uma determinada covariável muda ao longo do espaço. Neste cenário, recomenda-se o uso de modelos para dados de área, permitindo que o efeitos das covariáveis variem espacialmente. Neste trabalho, foi realizada a modelagem do número de casos registrados por COVID-19 nos 92 municípios que compõem o Estado do Rio de Janeiro, considerado o período entre o começo da pandemia e o fim do mês de janeiro de 2021, usando como variáveis explicativas as três dimensões que compõem o Índice de Vulnerabilidade Social (IVS) proposto pelo Instituto de Pesquisa e Economia Aplicada. Constatou-se que o modelo que permite que o efeito da dimensão de infraestrutura do IVS varie espacialmente foi o que melhor se ajustou aos dados.

Palavras-chave: CAR Intrínseco. COVID-19. Inferência Bayesiana.

Introdução

Em um contexto de dados espaciais, dados de área são aqueles que representam quantidades agregadas de uma variável de interesse em sub-regiões de uma região de estudo [2]. Dados como o número de casos de uma doença nos estados de um país ou de um determinado tipo de crime nos bairros de uma cidade são exemplos típicos.

Neste contexto, modelos que assumem uma distribuição de Poisson para as contagens observadas com a presença de efeitos aleatórios espacialmente estruturados são muito utilizados. A modelagem destes dados considerando este componente estruturado espacial permite avaliar a forma com que a organização espacial destas sub-regiões influenciam no fenômeno observado.

Porém existem cenários na qual a abordagem mais tradicional possa não ser a mais indicada. Por exemplo, cenários que apresentam alta variabilidade espacial ou situações problemas nas quais o pesquisador esteja interessado em entender se o impacto de uma determinada covariável é diferente em sub-regiões distintas de um espaço geográfico de interesse. Nestes cenários apresentados, o uso de um modelo com coeficientes variando no espaço pode ser mais adequado para lidar com a demasiada heterogeneidade espacial presente nos dados ou mesmo para ser possível responder os questionamentos levantados pelo pesquisador. Assim, a utilização de um conjunto adicional de parâmetros variando no espaço permitiria capturar toda a variação causada pela existência de uma estrutura espacial entre as sub-regiões. A abordagem utilizada neste trabalho para permitir que os coeficientes variem no espaço será a utilização de distribuições *a priori* do tipo condicionais autoregressivas para os efeitos aleatórios, de forma similar ao feito com os efeitos espaciais.

Com o objetivo de avaliar diferentes especificações de modelos com coeficientes variando no espaço, uma aplicação em dados de COVID-19 nos municípios do Rio de Janeiro foi realizada. Sendo definidas como covariáveis as três dimensões do Índice de Vulnerabilidade Social. O ajuste dos modelos foi realizado sob a perspectiva Bayesiana.

Material e métodos

O presente estudo consistiu na modelagem espacial dos casos de COVID-19 nos 92 municípios que compõem o Estado do Rio de Janeiro. A pandemia no Estado se desenvolveu de modo que nos primeiros meses os casos estivessem concentrados nos municípios da Região Metropolitana e, com o passar do tempo, ocorreu um processo de interiorização da mesma no qual os focos passaram a ser municípios do interior do Estado [4]. Este comportamento espaço-temporal gerou uma forte heterogeneidade espacial no número de casos fazendo com que seja um cenário propício para utilização de modelos com coeficientes variando no espaço.

O primeiro caso de COVID-19 no Rio de Janeiro foi confirmado no município de Barra Mansa em 05 de março de 2020, o primeiro óbito seria confirmado 14 dias depois no dia 19 de março no município de Miguel Pereira [6, 3]. Para este estudo de caso foi considerado o período entre o começo da pandemia e o fim do mês de janeiro de 2021. A delimitação deste período foi feita a partir de uma análise do comportamento da pandemia. Até então o número de casos se mantinha relativamente constante, porém após janeiro de 2021 houve um forte aumento no número de casos e que implicaria na necessidade do modelo acomodar um conjunto de efeitos temporais para capturar esta variação - o que não é o foco do estudo.

Os fatores envolvidos com o desenvolvimento da pandemia são complexos e o acesso a variáveis que pudessem explicá-lo nem sempre é trivial. Neste contexto optou-se por utilizar as três dimensões que compõem o Índice de Vulnerabilidade Social (IVS) proposto pelo Instituto de Pesquisa e Economia Aplicada. O IVS é um índice composto por três dimensões distintas: infraestrutura urbana, capital humano e renda e trabalho, cada qual sendo constituído por diversos indicadores calculados a partir de dados do Censo demográfico [5]. Espera-se que a utilização de suas dimensões separadamente permita avaliar o efeito de cada um dos aspectos que as mesmas representam no número de casos observados no Estado.

A classe de modelos escolhida para utilização neste estudo aplicado foi a dos modelos com distribuições condicionais autoregressivas para os efeitos espaciais. Assim, este estudo aplicado propõe uma comparação entre diferentes modelos: um deles utilizando um conjunto de efeitos espaciais e efeitos fixos para as covariáveis e um que permitia que que os coeficientes associados às covariáveis variassem espacialmente.

No caso em que o coeficiente varia no espaço, ao coeficiente é atribuída a distribuição proposta por Besag *et al.* (1991) que define o CAR Intrínseco [1]. Esta última distribuição também é atribuída ao conjunto de efeitos espaciais $\boldsymbol{\phi}$.

Neste trabalho, foram investigados dois modelos inicialmente. O primeiro assume fixo no espaço os efeitos de todas as covariáveis (Modelo 2) e o segundo assume que todos os efeitos variam espacialmente (Modelo 1). Após a avaliação dos resultados iniciais verificou-se que as estimativas de coeficientes espaciais para as dimensões de renda e capital humano eram bastante similares entre si. Deste modo, um terceiro modelo foi ajustado, no qual apenas a dimensão de infraestrutura possuía coeficientes variando no espaço (Modelo 3).

Para a definição do modelo final, assuma que Y_i e Pop_i são o número de casos por COVID-19 registrados e o tamanho da população no município i, $i = 1, \ldots, 92$, respectivamente. Considere também que foi observado o seguinte vetor de covariáveis $\boldsymbol{IVS}_i = (IVS_i^I, IVS_i^C, IVS_i^R)^T$, $i = 1, \ldots, 92$. O modelo será especificado como a seguir:

$$Y_i \sim Poisson(Pop_i \lambda_i),\ i = 1, \ldots, 92, \tag{1}$$

$$log(\lambda_i) = \alpha + \boldsymbol{IVS}_i^T \boldsymbol{\beta} + \phi_i,\ i = 1, \ldots, 92, \tag{2}$$

$$\beta_{1,i} \sim N\left(\frac{\sum_{j\sim i} w_{ij}\phi_j}{\sum_{j\sim i} w_{ij}}, \frac{1}{\tau_{\beta_1} \sum_{j\sim i} w_{ij}}\right), \tag{3}$$

$$\phi_i | \Phi_{-i}, W, \tau \sim N\left(\frac{\sum_{j\sim i} w_{ij}\phi_j}{\sum_{j\sim i} w_{ij}}, \frac{1}{\tau \sum_{j\sim i} w_{ij}}\right). \tag{4}$$

Para completar a especificação do modelo assumiu-se $\alpha \sim N(0, 0.001)$, $\beta_l \sim N(0, 0.0001), l = 2, 3$, $\tau \sim Gama(1, 0.05)$, $\tau_{\beta_1} \sim Gama(1, 0.05)$. No primeiro nível, é assumido uma distribuição Poisson para as contagens com média $Pop_i\lambda_i$. Pop_i funciona como um *offset*. É proposto um modelo log-linear para λ_i. Foram consideradas as três dimensões do IVS como covariáveis e um efeito espacial. Para o efeito da dimensão infraestrutura do IVS e o efeito espacial foi utilizada

uma distribuição *a priori* CAR Intrínseco e para as demais, foram utilizadas distribuições *a priori* relativamente vagas.

Por fim, a comparação entre os três modelos ajustados se deu a partir do *Watanabe-Akaike Information Criterion*, uma medida de qualidade de ajuste no qual o modelo que apresenta o valor mínimo é aquele que melhor se ajustou aos dados[7]. Ele é definido a seguir:

$$\begin{array}{rcl} WAIC & = & T_n + \frac{V_n}{n} \\ T_n & = & -\frac{1}{n}\sum_{i=1}^{n} logp*(X_i|\theta) \\ V_n & = & \sum_{i=1}^{n}\{E_\theta[(logp*(X_i|\theta))^2] - E_\theta[logP(X_i|\theta)]^2\} \end{array} \quad (5)$$

em que $p^*(\boldsymbol{x}|\theta)$ é a distribuição preditiva, T_n é a distribuição preditiva completa e V_n é a variância individual dos parâmetros na distribuição preditiva somada nas n observações. Este último termo cumpre o papel de ser um parâmetro de penalização de complexidade. Isso faz com que modelos mais parcimoniosos, ou seja, com menos parâmetros, sejam menos penalizados do modelos mais complexos.

Resultados e discussão

O período considerado para o estudo de caso foi do início da pandemia no Brasil, no começo de março de 2020, até o momento em que o número de casos explodiu, aproximadamente no final de janeiro de 2021. Neste intervalo de tempo os municípios do Estado do Rio de Janeiro confirmaram ao todo 523.414 casos de COVID-19, observando a nível municipal o Rio de Janeiro foi o município que registrou o maior número de casos: 187.281, representando 35,78% de todos os casos confirmados no Estado. Alternativamente, o município de Rio das Flores foi aquele com menor número de casos confirmados com apenas 45, ou seja, aproximadamente 0,008% dos casos confirmados.

O número médio de casos no Estado foi 5.689 com desvio padrão de 19.854 casos. Deve-se levar em conta que a pandemia se desenvolveu sob dinâmicas diferentes dependendo da região do Estado e também que diferentes municípios apresentavam capacidades diferentes de testar suas respectivas populações e identificar corretamente os casos de COVID-19. A Figura 1 apresenta a distribuição espacial dos casos de COVID-19 nos municípios do Rio de Janeiro.

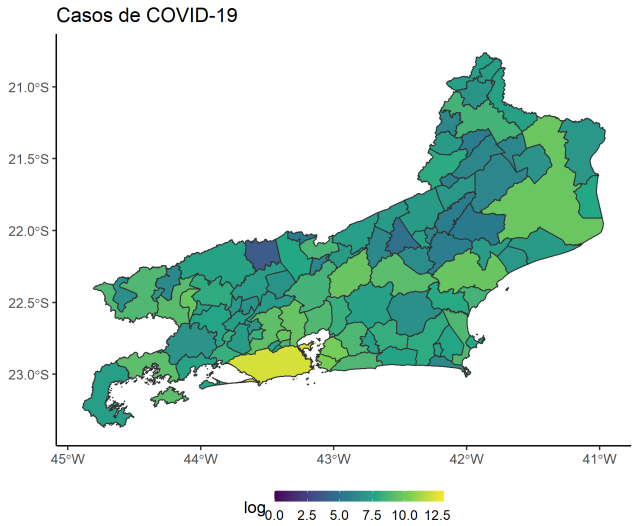


Figura 1: Distribuição espacial dos casos de COVID-19 no Estado do Rio de Janeiro.

Nota-se que os municípios da Região Metropolitana estão entre aqueles com maior número de casos confirmados. Além destes destacam-se também os municípios de Angra dos Reis, mais ao Sul do Estado, e Campos dos Goytacazes ao Norte entre aqueles com quantidades elevadas de casos de COVID-19. A análise do mapa coroplético permite estabelecer uma diferença no comportamento da pandemia entre os municípios mais próximos da costa e aqueles mais ao interior: quanto mais ao interior menor parecem ser os números de casos confirmados.

Por fim, uma comparação dos modelos foi feita por meio de uma medida de qualidade de ajuste, neste caso o WAIC. E notou-se que os modelos com coeficientes espaciais se adequavam melhor aos dados do que os modelos utilizando apenas efeitos fixos.

As estimativas pontuais e os respectivos intervalos de credibilidade de 95% estimados por cada modelo são apresentados a seguir na Tabela 1.

	Modelo 1		Modelo 2		Modelo 3	
	Mediana	$IC_{95\%}$	Mediana	$IC_{95\%}$	Mediana	$IC_{95\%}$
Intercepto	-3,439	[-3,660;-3,208]	-3,361	[-3,368;-3,355]	-3,315	[-3,429;-3,212]
Infraestrutura	-	-	-0,231	[-0,655;0,054]	-	-
Renda	-	-	0,175	[-0,115;0,593]	-0,015	[-0,175;0,182]
Capital Humano	-	-	-0,306	[-0,591;-0,059]	-0,143	[-0,299;0,046]
Precisão	0,450	[0,305;0,668]	0,397	[0,283;0,532]	1,165	[0,775;1,819]
Precisão (Infraestrutura)	3,068	[1,077;8,057]	-	-	3,029	[1,182;8,325]
Precisão (Renda)	3,802	[1,044;9,692]	-	-	-	-
Precisão (Capital Humano)	5,326	[2,153;12,095]	-	-	-	-
WAIC	1021,252	-	1044,806	-	1029,611	-

Tabela 1: Estimativas pontuais e intervalos de credibilidade de 95% para os parâmetros dos modelos ajustados. Para os modelos com coeficientes variando no espaço a estimativa do efeito fixo foi omitida.

As estimativas para o intercepto feitas pelos três modelos foram bastante similares entre si. Para o parâmetro de precisão as estimativas dos modelos 1 e 2 foram próximas, uma precisão baixa sugere a presença de maior variabilidade nas estimativas dos efeitos espaciais. Aqui deve-se notar também que a precisão estimada pelo modelo 3 foi razoavelmente mais elevada, o que era de se esperar, uma vez que como existem mais de um efeito capturando a heterogeneidade espacial, isso permite que os efeitos aleatórios não apresentem grande variabilidade.

A comparação dos modelos com base no WAIC sugeriria que o melhor modelo foi o modelo 1, porém as estimativas dos coeficientes variando no espaço das dimensões de capital humano e renda eram bastante similares entre si. Desta forma o modelo 3 foi escolhido como modelo final, ele trata alguns dos coeficientes como efeitos fixos e permite que o coeficiente da dimensão de infraestrutura varie no espaço. Em todo caso, os resultados do WAIC dos três modelos demonstram que a adoção de uma estratégia que permitisse a variação espacial dos coeficientes foi benéfica para a qualidade de ajuste: o modelo que utilizou apenas efeitos fixos para as variáveis foi o com pior desempenho entre os três.

Utilizando o modelo 3 como base de interpretação das estimativas dos coeficientes têm-se que quanto maior a vulnerabilidade de renda e capital humano há uma diminuição no número de casos observados. Entretanto, deve-se lembrar que municípios mais vulneráveis possivelmente possuem menor capacidade de testagem e assim, maior número de casos subnotificados. Por esta lógica, municípios menos vulneráveis tendem a possuir menor proporção de subnotificação e mais casos confirmados.

Uma análise cuidadosa da Figura 2 mostra a uma forte heterogeneidade das estimativas dos efeitos espaciais. Esta variabilidade nos coeficientes demonstrou que permitir que este coeficiente variasse no espaço era uma estratégia benéfica para o modelo. Quanto aos efeitos espaciais estimados, para a maior parte dos municípios, os intervalos de credibilidade não contém o zero, sendo observados também alguns municípios com intervalos substancialmente maiores, sugerindo maior incerteza sobre os mesmos. Já para o efeito do IVS, apesar de não apresentar uma grande variabilidade espacial, o modelo apresentou melhor comportamento de convergência com este efeito variando no espaço, assim como um melhor WAIC quando comparado com o modelo de efeitos fixos.

O estudo aplicado permitiu explorar um cenário no qual a utilização de um modelo com coeficientes variando no espaço foi benéfica para a qualidade de ajuste. Verificou-se também que uma análise exploratória sobre as covariáveis auxiliou no processo de decisão para determinar quais covariáveis apresentavam heterogeneidade espacial o suficiente para que seus coeficientes fossem tratados como efeitos aleatórios no espaço. Desta forma, conclui-se que, apesar da complexidade e custos computacionais adicionais, modelos com coeficientes variando no espaço são uma adição útil ao inventário de modelos espaciais e que podem ser mais vantajosos que modelos mais simples em um determinado número de cenários.

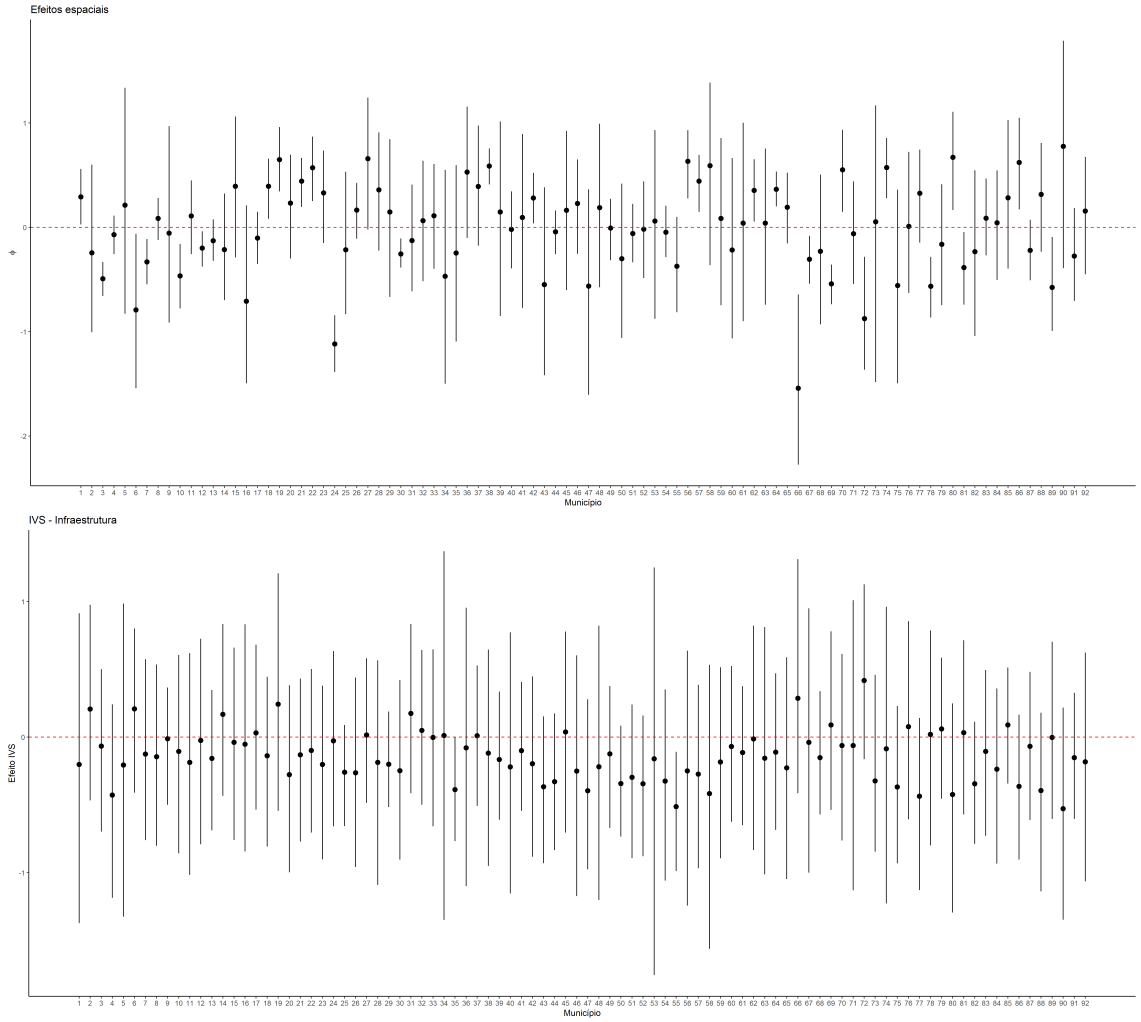


Figura 2: Estimativa pontual e intervalar para os efeitos espaciais e efeito da dimensão de infraestrutura do IVS para os 92 municípios do Estado.

Referências

[1] Besag, J. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B. 36* (1974), 192–236.

[2] Bivand, R. S., Pebesma, E. J. e Gomez-Rubio, V. *Applied spatial data analysis with R*, 2 ed. Springer, Nova Iorque, 2013.

[3] DATASUS. *Informações da Saúde (TABNET) - DATASUS*, 2021.

[4] de Souza, R. J., da Silva, P. V. e Junior, J. A. P. Modelando óbitos por covid-19 em cenários de heterogeneidade espacial: Uma abordagem utilizando modelos hierárquicos bayesianos. *Revista Brasileira de Estatística 78*, 245 (2020), 69 – 90.

[5] Instituto de Pesquisa e Economia Aplicada. *Atlas do Desenvolvimento Humano no Brasil*, 2020.

[6] Ministério da Saúde. *Painel Coronavírus*, 2021.

[7] Watanabe, S. A widely applicable bayesian information criterion. *Journal of Machine Learning Research 14* (2013), 867–897.

Identificação de clusters de roubos de veículos ocorridos na cidade do Rio de Janeiro entre 2016 e 2020

Filipe Nascimento (UFF)
Wu Xin (UFF)
Pedro Fernando (UFF)
Ricardo Junqueira (ISP)
Rafael Erbisti (UFF)
Jony Arrais Pinto Junior (UFF)

Email de contato: filipe_silva@id.uff.br.

Resumo

O município do Rio de Janeiro é o responsável por 21,21% dos roubos de veículo ocorridos entre 2016 e 2020 no Estado. A análise espacial deste tipo de delito permite identificar a existência de padrões espaciais nestas ocorrências e se existem agrupamentos de regiões com comportamento similar. Os resultados obtidos neste trabalho sugerem a existência de um agrupamento persistente na Zona Norte da cidade que anualmente apresenta números elevados deste tipo de delito para todos os anos avaliados.

Palavras-chave: Dados de área. Roubo de veículos. Análise de clusters. I de Moran.

1 Introdução

As frequentes notícias sobre criminalidade na cidade do Rio de Janeiro fizeram com que a mesma ganhasse a fama de ser uma das mais perigosas do país. Entretanto, esta fama não está tão distante da verdade. Dados do Atlas da Violência de 2019 mostram que o Rio de Janeiro é a cidade com maior número de homicídios entre aquelas com mais de 100 mil habitantes, ainda que sua taxa e homicídios sequer figure entre as cinquenta maiores do país [7].

Outro delito que recebe bastante atenção pelos números elevados no município é o roubo de veículos. Dados do Instituto de Segurança Pública do Rio de Janeiro mostram que de 2016 a 2020 o município do Rio de Janeiro registrou 100.805 roubo de veículos, representando 21,21% do total de registros deste delito para todo o Estado. A investigação da existência de padrão espacial nos crimes ocorridos em uma região já foi tema de estudo de diversos autores. Campos *et al.* (2019) utilizaram os índices de Moran Global e Local para investigar a dependência espacial nas ocorrências de roubos e furtos de veículos nos municípios do Rio Grande do Norte em 2017 [5]. Carrets *et al.* (2018) avaliou a existência de dependência espacial em crimes contra pessoa e patrimônio nos anos de 2005, 2010 e 2015 [6].

O presente estudo tem como objetivo investigar a dinâmica da distribuição espacial dos roubos de veículo no município do Rio de Janeiro entre os anos de 2016 e 2020. Este trabalho seguirá uma linha similar aos trabalhos citados, utilizando estatísticas como os índices de Moran para investigar a dinâmica espacial dos roubos de veículo no município. A unidade espacial escolhida para a análise foi a Circunscrição Integrada de Segurança Pública (CISP), uma subdivisão da área sob responsabilidade de um batalhão da Polícia Militar e que é equivalente a uma circunscrição de uma delegacia de Polícia Civil.

A partir da análise espacial deseja-se identificar *clusters* com dependência espacial significativa. Identificando clusters com altas taxas de roubos de veículos e avaliar se os mesmso se modificam ao longo do tempo.

2 Materiais e Métodos

2.1 Dados

Os dados de roubos de veículo utilizados neste estudo foram disponibilizados pelo Instituto de Segurança Pública do Rio de Janeiro (ISP-RJ). As contagens de roubos foram agregadas para as 41 CISPs que compõem o município do Rio de Janeiro para os anos 2016 a 2020. Por fim, foram utilizadas as populações anuais de cada CISP para calcular a taxa de roubo de veículos, de modo a eliminar o efeito da população nestas contagens.

2.2 Metodologia

A análise exploratória de dados espaciais consiste na utilização de métodos estatísticos para investigar a existência de comportamento espacial estruturado tanto a nível global quanto local. Entre estes métodos estão a construção de mapas coropléticos para visualizar a distribuição espacial da variável interesse quanto a utilização de indicadores que permitem verificar a existência da dependência espacial como também caracterizá-la [4].

O primeiro passo para avaliar a existência de dependência espacial é a construção de uma matriz de vizinhança que represente a estrutura espacial entre as unidades de área. Neste trabalho foi adotado um critério de contiguidade, em que duas unidades de área serão consideradas vizinhas caso as mesmas partilhem fronteiras.

A matriz de vizinhança W é definida da seguinte forma:

$$W = \begin{cases} w_{ij} > 0, & \text{se } i \sim j \\ w_{ij} = 0, & \text{c.c.} \end{cases}, \tag{1}$$

em que w_{ij} é o elemento da matriz de vizinhança W que indica se as sub-regiões i e j são vizinhas. A notação $i \sim j$ indica a relação de vizinhança.

Uma das medidas de identificação de autocorrelação espacial bastante populares é o Índice de Moran Global. Este é uma relação de autocovariância do tipo produto cruzado pela variância dos dados. Sua interpretação é feita da seguinte forma: se o valor da estatística for igual ao valor esperado, dentro dos limites de significância estatística, então não há autocorrelação espacial. Se o valor da estatística excede o valor esperado então há autocorrelação espacial positiva, ou seja, existe similaridade entre os valores observados e sua localização espacial [3]. Caso o valor da estatística seja menor que o valor esperado, então existe dissimilaridade entre o valor da variável e sua localização espacial.

O Índice de Moran Global é definido da seguinte maneira:

$$I = \frac{n}{S_0} \cdot \frac{\sum_i \sum_j w_{ij} \cdot z_i \cdot z_j}{\sum_{i=1}^{n} z_i^2} \tag{2}$$

em que n é o número de sub-regiões, z_i e z_j são os valores padronizados da variável de interesse nas sub-regiões i e j e S_0 é o somatório de todos os w_{ij}, ou seja, na especificação binária para W ele representa o número de vizinhos que a sub-região i possui.

O diagrama de dispersão de Moran possibilita a visualização e caracterização da autocorrelação espacial. A partir de um diagrama dividido em 4 quadrantes no qual cada um representa um tipo de associação linear espacial. O primeiro deles é denominado Alto-Alto (AA) e contém regiões com valores altos (acima da média) da variável interesse cujos vizinhos também apresentam valores altos. O quadrante Baixo-Baixo (BB) é o posto do AA e contém as regiões de associação espacial com valores baixos e cujos vizinhos apresentam valores baixos. Os outros dois quadrantes apresentam categorias mistas: o quadrante Baixo-Alto (BA) contém regiões com valores baixos referentes a variável de interesse, porém vizinhas de regiões de valores altos. Por fim, o quadrante Alto-Baixo (AB) contém regiões com valores altos cujos vizinhos possuem valores baixos da variável de interesse.

Para cenários espacialmente complexos é possível a ocorrência de autocorrelação espacial local, ou seja, entre pequenos grupos de regiões espacialmente dependentes entre si dentro da região de estudo. Uma possibilidade de investigar a existência de autocorrelação espacial local é a utilização do Índice de Moran Local, também chamado de Indicador de Associal Espacial Local (LISA) [2]. A partir dele é possível identificar *clusters* espaciais estatisticamente significativos para a variável de interesse, e o seu somatório é proporcional ao seu indicador de autocorrelação espacial global.

A estatística LISA é definida para cada região em função do valor da média dos seus vizinhos, possui as mesmas 4 classificações do diagrama de dispersão de Moran, porém a comparação é feita com o valor médio das regiões vizinhas. Ele é definido da seguinte forma:

$$I_i = \frac{\sum w_{ij} \cdot z_i \cdot z_j}{\sum_{i=1}^{n} z_i^2} \tag{3}$$

em que z_i e z_j são os valores padronizados da variável de interesse nas sub-regiões i e j e S_0 é o somatório de todos os w_{ij}, ou seja, na especificação binária para W ele representa o número de vizinhos que a sub-região i possui.

A partir da utilização de mapas coropléticos e destas três estatísticas foi possível realizar uma análise exploratória completa para determinar a existência de dependência espacial significativa a níveis global ou local. Os resultados desta análise serão apresentados na próxima Seção.

3 Resultados e discussão

No município do Rio de Janeiro os roubos de veículo representaram 19% do total deste tipo de delito registrado entre 2016 e 2020. As tabelas 1 e 2 a seguir apresentaram as 5 CISPs que registraram os maiores e menores números de roubos de veículos no período analisado.

CISP	Ocorrências	Min.	1° Quartil	Média	Mediana	3° Quartil	Max.
39	9189	1243	1519	1838	1750	1947	2730
34	8541	1169	1347	1708	1652	2176	2197
27	6654	582	1382	1331	1450	1523	1717
40	6599	412	1021	1320	1477	1515	2174
31	6143	905	1115	1229	1300	1386	1437

Tabela 1: Medidas descritivas das 5 CISPs com maior número de ocorrências no período analisado.

CISP	Ocorrências	Min.	1° Quartil	Média	Mediana	3° Quartil	Max.
11	10	1	2	2	2	2	3
13	46	6	9	9.2	9	10	12
12	51	3	9	10.2	11	14	14
1	102	4	12	20.4	20	27	39
14	148	22	26	29.6	29	32	39

Tabela 2: Medidas descritivas das 5 CISPs com menor número de ocorrências no período analisado.

Nota-se aqui que entre as cinco CISPs que apresentaram o maior número de ocorrências de roubos de veículos no período quatro delas fazem parte da Zona Norte do Rio de Janeiro, uma região notória pelo número de favelas e pelas elevadas taxas de criminalidade. Alternativamente, todas as CISPs que apresentaram os menores números de roubos de veículo fazem parte da Zona Sul - a região mais nobre do município. Estas estatísticas descritivas já sugerem parte dos resultados que serão vistos na análise: pode-se esperar *clusters* de números elevados de roubos de veículos na Zona Norte e *clusters* de números baixos na Zona Sul.

A distribuição espacial dos roubos de veículos nos anos 2016 a 2020 é apresentada a seguir na Figura 1.

Nota-se que ao longo dos cinco anos analisados houve um certo padrão no comportamento espacial dos roubos de veículo. Em geral, os números elevados de roubos estavam concentrados na Zona Norte do município, se concentrando em torno dos bairros de Realengo, Madureira, Pavuna, Irajá e Anchieta. Nota-se também uma forte redução em todo o município no ano de 2020, esta é facilmente explicada pela pandemia da COVID-19 que gerou restrições de movimentação e consequentemente uma queda no número de delitos.

O passo seguinte foi avaliar a existência de autocorrelação espacial global. A Tabela 3 a seguir apresenta os resultados do Índice de Moran Global.

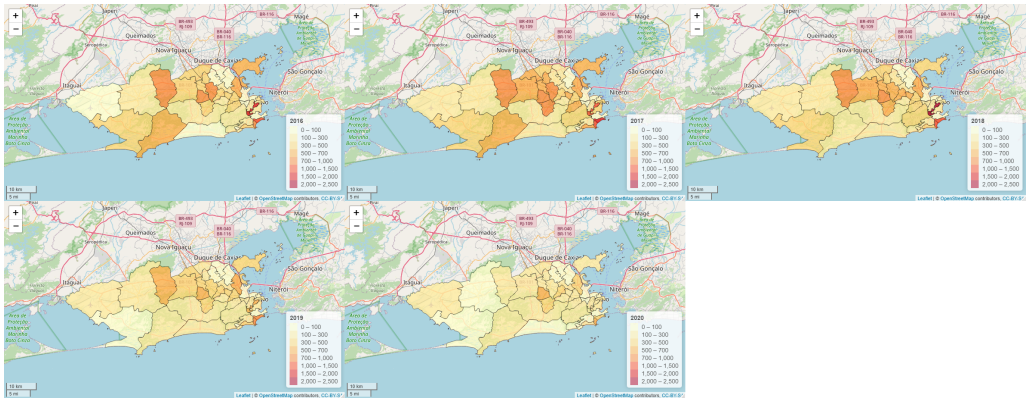


Figura 1: Sub-regiões espacialmente dependentes pelo Índice de Moran Local.

Ano	I de Moran	Esperado	p-Valor
2016	0.4311	-0.025	< 0.0001
2017	0.4872	-0.025	< 0.0001
2018	0.3056	-0.025	0.0001
2019	0.5496	-0.025	< 0.0001
2020	0.5657	-0.025	< 0.0001

Tabela 3: Teste de autocorrelação espacial de Moran.

Avaliando pelo p-valor, sob um nível de significância de 5%, tem-se que para todos os anos foi observada dependência espacial global significativa. Nota-se também que um valor positivo para o Índice de Moran Global sugere que a autocorrelação espacial é positiva, ou seja, números de roubo de veículos similares tendem a se agrupar na região de estudo.

Estes resultados são corroborados pelos resultados do diagrama de dispersão de Moran, conforme apresentado na Figura 2 a seguir.

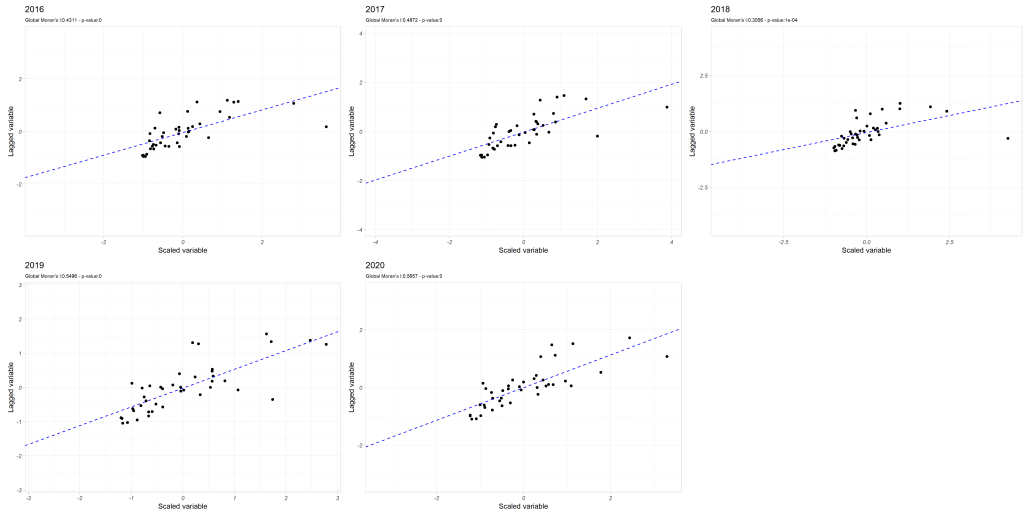


Figura 2: Diagramas de dispersão de Moran.

Embora grande parte das observações esteja concentrada perto do centro dos eixos, nota-se que há um número significativo de observações no quadrante referente a categoria 'Alto-Alto'. Este resultado sugere autocorrelação espacial positiva e também que estas sub-regiões possuem valores elevados e com vizinhos de valores similares.

Este resultado ficará mais facilmente interpretável com a apresentação dos resultados do Índice de Moran Local na Figura 3 a seguir.

Para todos os anos analisados foram encontrados *clusters* de sub-regiões com dependência espacial tanto positiva quanto negativa. Uma análise rápida permite perceber que para todo o

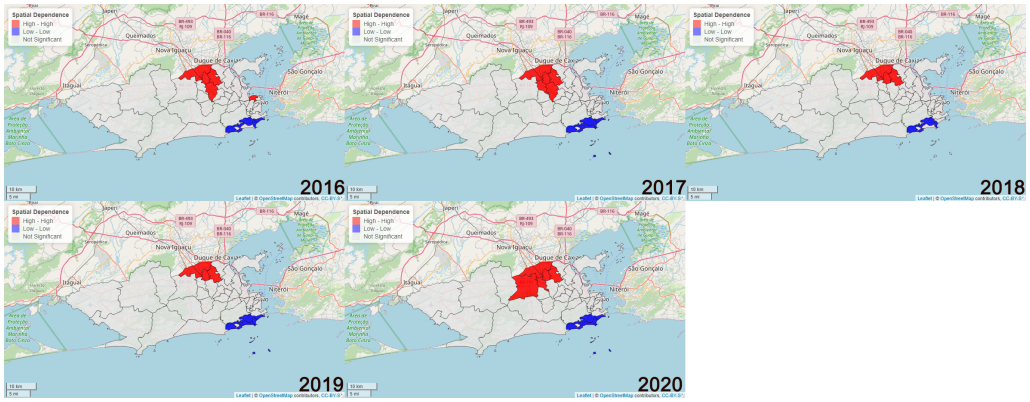


Figura 3: Clusters de sub-regiões espacialmente dependentes pelo Índice de Moran Local.

período o comportamento espacial dos *clusters* foi quase constante: existe sempre um *cluster* da categoria Alto - Alto na Zona Norte e sempre um Baixo - Baixo na Zona Sul, havendo apenas algum grau de variabilidade sobre quais CISPs fazem parte dos mesmos a cada ano.

Ainda que exista esta variabilidade, deve-se levar em consideração as características das CISPs em ambos os *clusters*. No caso do *cluster* da Zona Norte deve-se perceber que esta região abriga vias importantes e que apresentam grande fluxo de veículos como a Avenida Brasil e a Rodovia Presidente Dutra. Esta região também possui rendas menores, maiores taxas de criminalidade e uma quantidade elevada de favelas. O *cluster* da Zona Sul contém as CISPs que fazem parte de alguns dos bairros de renda mais elevada do Rio de Janeiro, isto acarreta em uma série de fatores que impactam na ocorrência deste tipo de delito na região.

As ferramentas utilizadas servem para fazer uma análise prévia da dinâmica espacial deste tipo de delito. Os resultados aqui apresentados sugerem que as CISPs que fazem parte do *cluster* na Zona Norte merecem mais atenção em relação ao roubo de veículos. De forma geral, esta é uma ferramenta útil para identificar padrões nas ocorrências que pode auxiliar no desenvolvimento de estratégias para reduzir o número de delitos. Como um desenvolvimento futuro está a possibilidade de utilizar modelos de identificação de *clusters* ou modelos de estratificação de risco, como aqueles propostos por Adin *et al.* (2019) e Lee & Mitchell (2014) [1, 8].

Referências

[1] Adin, A., Lee, D., Goicoa, T. e Ugarte, M. D. A two-stage approach to estimate spatial and spatio-temporal disease risks in the presence of local discontinuities and clusters. *Statistical Methods in Medical Research 28*, 9 (2018), 2595–2613.

[2] Anselin, L. Local indicators of spatial association - lisa. *Geographical Analysis 27*, 2 (1995), 93–115.

[3] Anselin, L. *An Introduction to Spatial Autocorrelation Analysis with GeoDa*, 2003.

[4] Bivand, R. S., Pebesma, E. J. e Gomez-Rubio, V. *Applied spatial data analysis with R*, 2 ed. Springer, Nova Iorque, 2013.

[5] Campos, J. e de Lima, L. C. Roubo e furto de veículos nos municípios do estado do Rio Grande do Norte em 2017.

[6] Carrets, F. D., de Oliveira, J. e Menezes, G. R. A criminalidade no Rio Grande do Sul: uma análise espacial para anos de 2005, 2010 e 2015.

[7] Instituto de Pesquisa Econômica Aplicada - IPEA. *Atlas da Violência*, 2019.

[8] Lee, D. e Mitchell, R. Boundary detection in disease mapping studies. *Biostatistics 13*, 3 (2012), 415–426.

Evolução temporal do desmatamento e de seus indicadores: um olhar para as Regiões Norte e Centro-Oeste do Brasil

Igor Da Silva Freitas De Souza (UFF)

Email de contato: igorfsouza18@gmail.com

Resumo

Ao longo dos últimos anos, muito tem-se falado sobre a existência de relação entre o aumento na área desmatada, especialmente da região amazônica, com o aumento da ocupação da mesma região por atividades agropecuárias e, também, pelo aumento sucessivo da quantidade de queimadas realizadas. A proposta do presente trabalho será, em primeiro lugar, modelar individualmente as séries temporais de quatro tipos de indicadores, que possibilitem a produção de previsões, bem como identificar se existe correlação temporal entre esses indicadores. Serão utilizados os dados disponibilizados pelos sistemas coordenados pelo Instituto Nacional de Pesquisas Espaciais (INPE), dados da pesquisa da pecuária municipal realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e dados sobre área desmatada e área destinada à agropecuária fornecidos pelo projeto MapBiomas.

Palavras-chave: Modelagem, Série Temporal, Box-Jenkins, Desmatamento, Agropecuária, Queimadas.

Introdução

Segundo a Food and Agriculture Organization (FAO) [4], 31% da superfície da terra é coberta por florestas, sendo a floresta tropical a de maior presença. O clima delas é quente e úmido, há longos períodos de chuvas e alguns períodos de seca [5].

As savanas possuem clima tropical, os períodos de secas podem durar até nove meses e a temperatura chega a 40. Cerca de 20% da população mundial vivem em áreas que são ou foram cobertas por savanas [15].

A Floresta Amazônica é a maior floresta tropical do mundo, e cerca de 60% dela se encontra no Brasil. Já o Cerrado é conhecido como a savana com a maior biodiversidade do mundo [10]. Esses dois Biomas ocupam quase toda a extensão territorial das Regiões Norte e Centro-Oeste do Brasil.

Em seu relatório, a FAO [4] afirma que desde 1990, o planeta já perdeu 420 milhões de hectares das duas florestas naturais. Porém, parece haver um decréscimo no desmatamento, que caiu de 16 milhões de hectares por ano entre 1990 e 2000, para 10 milhões de hectares por ano entre 2015 e 2020.

Há diversas regiões no globo que possuem alta taxa de desmatamento, mas em Ritchie (2020) [13] percebe-se que o Brasil é o país que possui a maior quantidade de desmatamento, possuindo uma média de mais de 5 milhões de hectares por ano, entre os anos de 2015 à 2020.

O presidente do Brasil, Jair Bolsonaro, afirmou na cúpula de líderes mundiais sobre o clima, que cumprirá os acordos assumidos no Acordo de Paris, que visam diminuir a emissão de gás carbônico. Para tal, Bolsonaro prometeu a eliminação do desmatamento até o ano de 2030 [1]. Porém, as medidas adotadas pelo governo foram no sentido oposto, obtendo nos anos de 2019 e 2020 tivemos as maiores quantidades de área desmatada desde 2008 [8]

Um estudo de Sy et al. (2015) [16] aponta que que 80% do desmatamento no Brasil é devido ao agronegócio, sendo a área de pastagem responsável pela maior parte, e a área de plantação responsável pelo restante.

Segundo Angelo e Sá (2007) [2] o efetivo de rebanho bovino é altamente significante para explicar o desflorestamento na Floresta Amazônica brasileira.

A queimada é uma prática utilizada para abrir áreas para plantações e pastos. Essas queimadas muitas vezes perdem controle e o fogo destrói áreas de floresta naturais [17].

O **objetivo** geral deste trabalho é construir modelos que possam descrever a evolução anual das séries temporais referentes à área de floresta natural, aos indicadores agro-pecuários, sendo esses a área destinada à lavouras e pastagens e o número de cabeças de gado bovino, e à quantidade de focos de queimadas nas regiões Norte e Centro-Oeste, além de verificar a existência de relação temporal entre elas.

Os **objetivos específicos**, aplicados a ambas as regiões, são apresentados abaixo:

- Modelar a série temporal da área coberta por florestas naturais
- Construir um modelo para a série temporal da quantidade de área destinadas à lavouras
- Modelar a série temporal da quantidade de área do solo coberta por pastagens
- Construir um modelo para a série temporal do efetivo de gado bovino
- Modelar a série temporal da quantidade de focos de queimadas
- Verificar se existe correlação cruzada entre as séries temporais investigadas
- Construir modelos para as séries temporalmente dependentes

Material e métodos

O Quadro 1 contêm a descrição das bases de dados

Bases de dados	Descrição
Floresta Nativa	Área de floresta nativa de 1985 à 2019
Pastagem	Área de pastagem entre 1985 e 2019
Efetivo bovino	Quantidade de cabeça de gado bovino de 1985 à 2019
Agricultura	Área destinada à lavouras entre 1985 e 2019
Queimadas	Quantidade de focos de queimadas de 1999 à 2020

Quadro 1: Bases de dados

As bases da Floresta Nativa, Pastagem e Agricultura foram coletadas do Projeto MapBiomas. Os dados são referentes aos agregados anuais das Regiões Norte e Centro-Oeste [11].

A base do Efetivo bovino foi coletado do Instituto Brasileiro de Geografia e Estatística (IBGE), e se refere ao agregado anual de todos os Estados das Regiões Norte e Centro-Oeste [7].

A base sobre Queimadas também é o agregado anual dos focos para as Regiões Norte e Centro-Oeste, e foi coletada do Projeto BDQueimadas do Instituto Nacional de Pesquisa Espaciais (INPE) [9].

Segundo Morettin e Toloi (2006) [12], uma série temporal é qualquer conjunto de observações que estejam ordenadas no tempo, e tem por objetivos descrever o comportamento da série, verificar a existência de periodicidades, investigar o mecanismo gerador da série temporal e realizar previsões de valores futuros.

O modelo clássico de uma série temporal, Z_t, é descrito da seguinte forma:

$$Z_t = T_t + S_t + a_t, \qquad \text{t} = 1, 2, \ldots, \text{n}.$$

Onde t é o tempo, T_t representa a tendência da série, S_t a sazonalidade e a_t é um fator aleatório. Neste trabalho o tempo t é em anos.

Modelos para descrever séries temporais, em geral, se baseiam em alguma suposição simplificadora. A mais comum entre elas é a de estacionariedade.

Um processo estocástico Z = {Z(t), t ∈ T} diz-se estritamente estacionário se todas as distribuições finito-dimensionais permanencem as mesmas sob translações no tempo, ou seja,

$$\text{F}(z_1,\ldots,z_n;t_1+\text{r},\ldots,t_n+\text{r})=\text{F}(z_1,\ldots,z_n;t_1,\ldots,t_n),$$

para quaisquer $t_1,\ldots,t_n$,r de T.

Um modelo paramétrico, que possui número de parâmetros finito, tenta descrever o comportamento de uma série temporal no domínio do tempo. Os modelos mais comum dessa classe são os modelos de regressão(linear ou de curva de crescimento), modelos auto-regressivos e de médias

móveis (ARMA), modelos auto-regressivos integrados e de médias móveis (ARIMA), modelos de memória longa (AFIMA), modelos estruturais e modelos não-lineares [12].

O modelo ARMA(p,q) tem o objetivo de ser parcimonioso, fazendo a junção dos modelos das classes auto regressivos, AR(p), e médias móveis, MA(q). Este modelo é uma ferramenta muito comum em análises de séries temporais, onde tem a finalidade de entender o comportamento da série ao longo do tempo e, talvez, prever valores futuros. Seja $\tilde{z}_t$ uma série estacionária

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

O modelo ARIMA(p,d,q) é uma generalização de um modelo ARMA(p,q), onde a série original $\tilde{z}_t$ é não estacionária, então ela é diferenciada d vezes até se obter uma série $\Delta^d \tilde{z}_t$ estacionária. Para casos onde $d \geq 1$, $\Delta^d \tilde{z}_t = \Delta^d z_t$, então, pode-se escrever o modelo da seguinte forma[3]

$$\Delta^d z_t = \phi_1 \Delta^d z_{t-1} + \cdots + \phi_p \Delta^d z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

Em uma série temporal sem componente da sazonalidade, a metodologia de Box-Jenkins aborda a previsão com base em modelos da classe auto-regressivo integrado e de médias móveis, ou, ARIMA(p,d,q). O método passa pelas etapas de identificação, estimação e validação do modelo. A partir da escolha do melhor modelo, pode-se avaliar a capacidade preditiva do modelo adotado [3].

A etapa de identificação é a fase crítica, ela visa determinar os valores de p, d e q do modelo ARIMA(p,d,q), esse processo é constituído por três partes[12]:

- Verificar a necessidade de transformação da série original como o intuito de estabilizar a variância. Tal identificação pode ser feita de forma descritiva, através dos gráficos dependentes do tempo
- diferenciar a série obtida no item anterior até obter uma série estacionária, a fim de reduzir o processo $\Delta^d Z_t$ a uma ARMA(p,q). O teste de raiz unitária será de grande utilidade para esta verificação
- Identificar valores p e q da ARMA(p,q) através das funções de autocorrelação e autocorrelação parcial

A etapa de estimação é aquela em que serão obtidas estimativa para os vários valores dos coeficientes $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$. Os estimadores para $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$ podem ser encontrados através de algumas formas, como pelo método dos momentos, método da máxima verossimilhança ou estimação não linear [12].

A etapa de validação consiste na verificação de independência e normalidade dos resíduos. Uma forma de verificar a independência dos resíduos é através do teste de Box-Pierce [12], e para verificação da normalidade pode-se usar o teste Shapiro-Wilk [14].

Vários procedimentos de identificação de modelos ARMA foram propostos, alguns consistem em funções penalizadoras. A ideia é escolher as ordens k e l que minimizem a quantidade

$$P(k,l) = \ln \hat{\sigma}^2_{k,l} + (k+l)\frac{C(N)}{N}$$

em que $\hat{\sigma}^2_{k,l}$ é a estimativa para a variância residual do modelo ARMA(k,l) ajustado às N observações da série temporal e C(N) é uma função do tamanho da série.

O termo penalizador $(k+l)\frac{C(N)}{N}$, aumenta quando novos parâmetros são adicionados, enquanto a variância residual diminui. Assim, o intuito é identificar as ordens k e l que equilibrem esse comportamento. Procedimentos de identificação que minimizam funções penalizadoras particulares são usualmente utilizadas para a escolha do melhor modelo.

O critério de informação de akaike(AIC) escolhe o melhor modelo cujas ordens k e l minimizem o seguinte critério

$$AIC(k,l) = \ln \hat{\sigma}^2_{k,l} + \frac{2(k+l)}{N}$$

O critério de informação Bayesiano(BIC) sugere que o melhor modelo é aquele que minimiza o seguinte critério

$$BIC(k,l) = \ln \hat{\sigma}^2_{k,l} + (k+l)\frac{\ln N}{N}$$

onde $\ln \hat{\sigma}^2_{k,l}$ é a estimativa de máxima verossimilhança da variância do modelo ARMA(k,l) [12].

Através de algumas métricas, pode-se verificar se o modelo representa bem o comportamento da série, ou seja, a diferença do valor efetivo e do valor ajustado pelo modelo é bem pequena.

As métricas comumente usada são o Erro Absoluto Médio (MAE), a Raiz do Erro Quadrático Médio (RMSE) e o Erro Médio Absoluto Percentual (MAPE) que tem a vantagem de ser em termos percentuais e, então, não atrelado a nenhuma unidade de medida [6].

Resultados e discussão

Os comportamentos das séries históricas são apresentados na Figura 1. A série temporal da Floresta Nativa parece ter comportamento inverso das demais, o que sugere que pode haver relação entre o desmatamento e os indicadores em estudo.

A série que difere de comportamento linear é a de Queimadas, onde não parece haver padrão. Porém, nos últimos anos parece estar em crescimento.

Devido aos comportamentos, ao decorrer dos anos, parecerem bem comportados, nenhuma transformação foi considerada para estabilização da variância.

Comportamentos cíclicos não foram identificados, o que sugere que as séries não possuem sazonalidade.

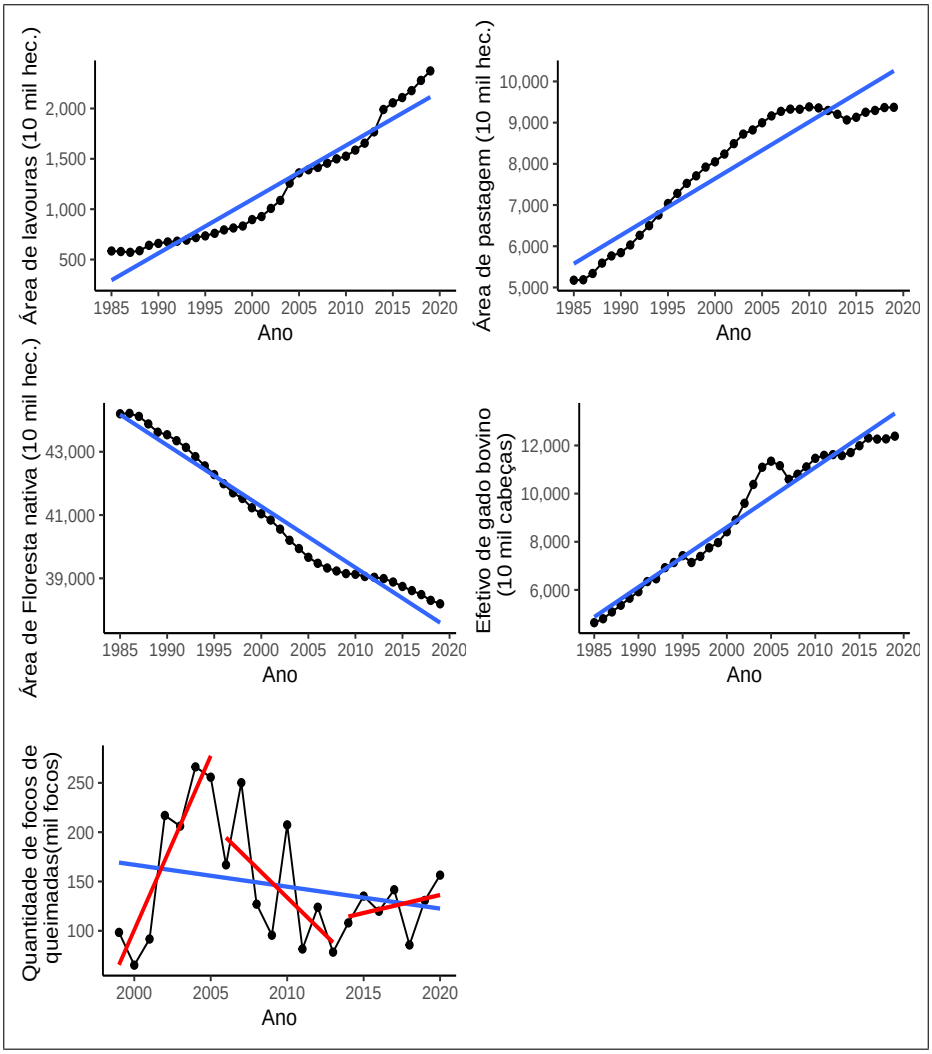


Figura 1: Séries temporais em estudo e suas tendências

Primeiramente foi verificado a estacionariedade das séries e foi identificado que nenhuma delas possuía, originalmente, comportamento estacionário. Então as séries foram diferenciadas até que a estacionariedade fosse estatisticamente aceita.

Possuindo as séries estacionárias, foi possível, através da comparação de modelos, identificar o melhor modelo ARIMA para cada uma das séries. Na Tabela 1 é possível ver os modelos escolhidos.

Tabela 1: Melhores modelos baseados nos critérios de informação para combinações de p e q ≤ 2 e métricas de qualidade de ajuste

Série	Melhor Modelo
Floresta Nativa	ARIMA(0,2,1)
Pastagem	ARIMA(2,2,0)
Efetivo Bovino	ARIMA(1,2,1)
Agricultura	ARIMA(1,1,0)
Focos de Queimadas	ARIMA(2,1,1)

A Figura 2 apresenta o ajuste dos modelos aos dados. É possível perceber que apenas o modelo para a série de Queimadas parece não se adequar bem.

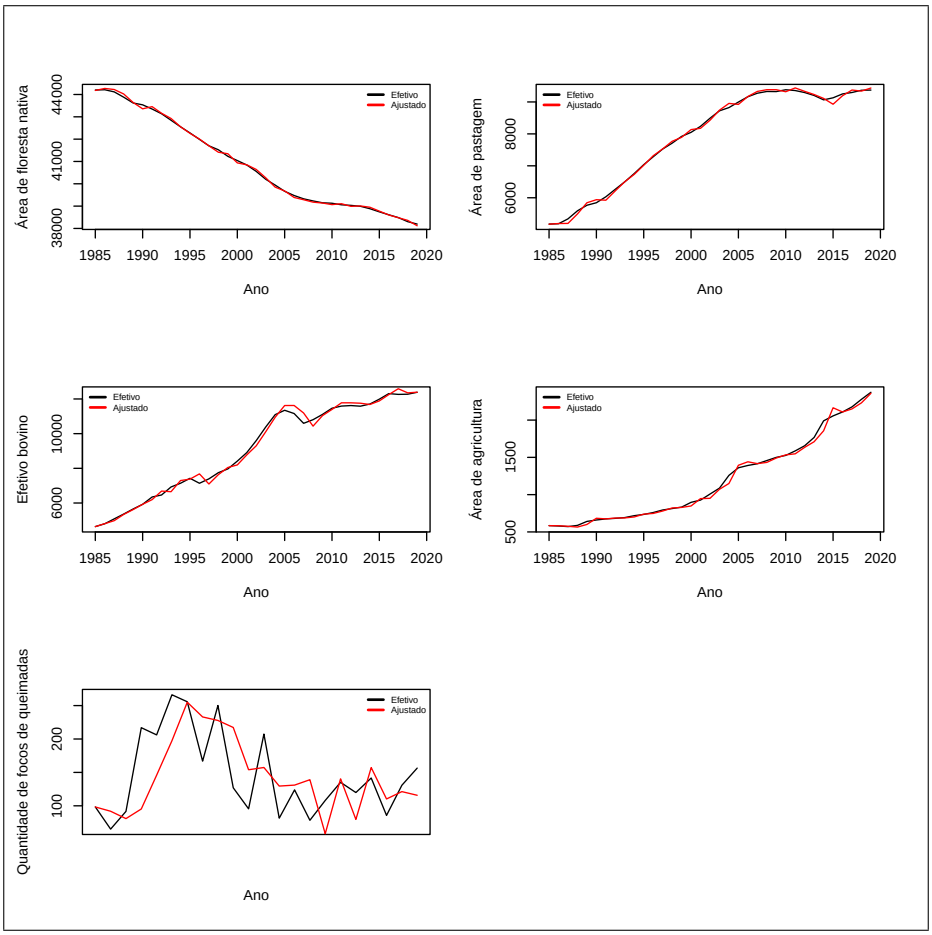


Figura 2: Valores efetivos e ajustados das séries temporais

A fim de validar o modelo, foi verificado a não correlação e a normalidade dos resíduos. A normalidade dos resíduos da série foi rejeitada, porém uma aproximação pela normal foi considerada.

O trabalho ainda não está concluído, os próximos passos são identificar outro modelo que possa se adequar melhor aos dados sobre focos de queimadas. Também irá se avaliar a dependência temporal das séries, e construir os modelos temporalmente dependentes, em caso afirmativo, além de realização de previsões utilizando esses modelos. A adição de uma nova base de dados sobre a quantidade de madeira extraída também está sendo considerada.

Referências

[1] Agência Brasil - Brasilia. Confira discurso do presidente bolsonaro na cúpula do clima, 2021. acesso em: 01 jul 2021.

[2] Angelo, H. e de Sá, S. P. P. O desflorestamento na amazônia brasileira. *Ciência Florestal* (2007).

[3] BOX, G. E. P. e JENKINS, G. M. *Time Series Analysis: Forecasting and Control*, revisada ed. Holden-Day, 1976.

[4] Food and Agriculture Organization. Global forest resources assessment 2020 – key findings. Acesso em: 22 jul 2021, 2020.

[5] Gallery, R. E. *Ecology of Tropical Rain Forests*. Ecology and the Environment, 2014, pp. 247–272.

[6] Hyndman, R. J. e Athanasopoulos, G. *Forecasting: principles and practice*, 2 ed. OTexts: Melbourne, Australia, 2018.

[7] IBGE. Pesquisa da pecuária municipal, 2020. Acesso em 22 ago 2021.

[8] INPE. Monitoramento do desmatamento da floresta amazônica brasileira por satélite, 2021. Acesso em: 22 ago 2021.

[9] INPE. Monitoramento dos focos ativos por região, 2021. Acesso em: 22 ago 2021.

[10] Klein, A. L., Ed. *Eugen Warming e o cerrado brasileiro: um século depois*. Editora UNESP, 2002.

[11] MAPBIOMAS. EstatÍsticas, 2020. Acesso em 22 ago 2021.

[12] Morettin, P. A. e de Castro Toloi, C. M. *Análise de séries temporais*, 2 ed. Edgard Blucher, 2006.

[13] Ritchie, H. Deforestation and forest loss, 2020. Acesso em: 01 jul 2021.

[14] Shapiro, S. S. e Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika 52*, 3/4 (1965), 591–611.

[15] Solbrig, O., Medina, E. e Silva, J. *Biodiversity and Savanna Ecosystem Processes: A Global Perspective*. John Wiley and Sons Ltd, 1996.

[16] Sy, V. D., Herold, M., Achard, F., Beuchle, R., Clevers, J. G. P. W., Lindquist, E. e Verchot, L. Land use patterns and related carbon losses following deforestation in south america.

[17] WWF-Brasil. Agricultura e pecuária, 2005. Acesso em: 6 jul 2021.

Doenças cardiovasculares e variáveis ambientais nos municípios da Amazônia Legal nos meses de seca de 2019

Isabelle de Oliveira Pinto (UFF)
Ludmilla da Silva Viana Jacobson (UFF)

Email de contato: isabellepinto@id.uff.br, ludmillajacobson@id.uff.br.

Resumo

O objetivo principal deste estudo foi avaliar as variáveis ambientais e internação por doenças cardiovasculares nos municípios da Amazônia Legal. A área de estudo é definida pelos municípios que compõem a Amazônia Legal e o período limitado ao ano de 2019, nos meses de Maio, Junho, Julho, Agosto, Setembro e Outubro. A escolha desses meses é devido a ocorrência de queimadas, quando geralmente se observa um aumento nas internações hospitalares. Foram utilizados os dados do SISAM, especificamente as concentrações de material particulado fino $PM_{2.5}$, temperatura e umidade relativa por município. Para o desfecho de saúde foram utilizados os dados disponíveis no DATASUS, por meio do Sistema de Internações Hospitalares do SUS (SIH/SUS). Na análise dos dados utilizamos o dendrograma e mapas para avaliar descritivamente a relação entre as variáveis e sua distribuição espacial. Todas as análises foram realizadas no programa estatístico R. Os resultados sugerem uma relação positiva entre as variáveis temperatura, $PM_{2.5}$ e taxa de internação e também relação inversa entre a umidade e as demais variáveis ambientais.

Palavras-chave: Amazônia Legal, Doenças Cardiovasculares, Poluição atmosférica, Queimadas, Concentração de $PM_{2.5}$.

Introdução

A Amazônia Legal é composta pelos Estados Acre, Amapá, Amazonas, Maranhão, Mato Grosso, Pará, Rondônia, Roraima e Tocantins e ocupa cerca de 58,9 % do território brasileiro. Neste trabalho foi investigada a associação entre poluição atmosférica, causada pelas queimadas, e internação por doenças cardiovasculares no ano de 2019 na Amazônia Legal. Foram escolhidos os meses de Maio à Outubro pois são os meses com menor índice de umidade. O ano de 2019 foi selecionado para este estudo devido ao aumento de aproximadamente 45,1% de focos de queimadas em relação ao ano anterior, 2018. No ano de 2019 os 6 Estados brasileiros com maior porcentagem de focos de queimadas pertencem à Amazônia Legal, sendo o primeiro lugar ocupado pelo Mato Grosso. [3] As doenças do aparelho circulatório foram escolhidas para este trabalho por ser a principal causa de morte em todo mundo nos últimos 20 anos. [6]

O $PM_{2.5}$, também conhecido como material particulado fino, têm se mostrado um indicador robusto de risco à saúde. O limite diário de exposição ao material particulado fino, em áreas urbanas, é de $25\mu g/m^3$ e o limite para a média anual é de $10\mu g/m^3$, de acordo com a OMS.[5] Estudos sugerem efeitos dos extremos de temperatura em diferentes desfechos de saúde, com destaque para as doenças cardiovasculares. [1]

As discussões neste trabalho visam explorar a distribuição espacial das variáveis ambientais $PM_{2.5}$, temperatura e umidade relativa, e sua relação com a taxa de internação por doenças cardiovasculares. Existem poucos estudos sobre o tema no ano de 2019, onde houve um aumento expressivo no número de focos de queimadas na Amazônia Legal. O objetivo geral deste trabalho foi avaliar descritivamente a relação entre as variáveis por meio de mapas e dendrograma.

Material e métodos

A área de estudo é composta pelos 772 municípios pertencentes à Amazônia Legal. Foram incluídos no estudo os indivíduos internados por doenças cardiovasculares residentes nos municípios da

Amazônia Legal no ano de 2019, nos meses Maio, Junho, Julho, Agosto, Setembro e Outubro, que fazem parte do período de seca. [2]

Os dados sobre as internações por doenças cardiovasculares foram obtidos através do site do Departamento de Informática do SUS (DATASUS), por meio do Sistema de Internações Hospitalares (SIH/SUS) e código CID-10 I00-I99, capítulo IX, Doenças do aparelho circulatório. A variável número total de internações por mês e por local de residência foi composta por 4632 observações. A partir da variável número de internações e da população estimada no ano de 2019 foi criada a variável de estudo Taxa de Internação por doenças cardiovasculares, cujo seu cálculo é feito pela divisão entre o número de internações e a população estimada, em cada município.

Os dados sobre umidade, temperatura e concentração de $PM_{2.5}$ foram obtidos no site do Sistema de Informações Ambientais Integrado a Saúde (SISAM). Os dados estavam na base original por dia e horário e a partir disso foram criadas novas variáveis agregadas como a média, mínimo, máximo, percentil 90 e amplitude, de cada mês por município. No total ficaram 13 variáveis ambientais.

Para verificar a proximidade entre as variáveis de pesquisa foi utilizado um diagrama chamado dendrograma. O método para construção do dendrograma foi o método de Ward, que é um procedimento de agrupamento hierárquico. [4]

Todas as análises e modificações no banco de dados foram feitas utilizando o programa RStudio. [7] Para a criação do dedrograma no RStudio foram utilizados os pacotes *ggdendro*, onde foi utilizada a função *ggdendrogram*, o pacote *dendextend*, onde foi utilizada a função *cutree*, e o pacote *cluster*, onde foram utilizadas as funções *agnes* e *pltree*. Além disso, também foram produzidos diversos mapas coropléticos onde foi utilizado a função *tm_ shape* do pacote *tmap*. [7]

Resultados e discussão

Para a análise exploratória foi feito um dendrograma que está na figura 1:

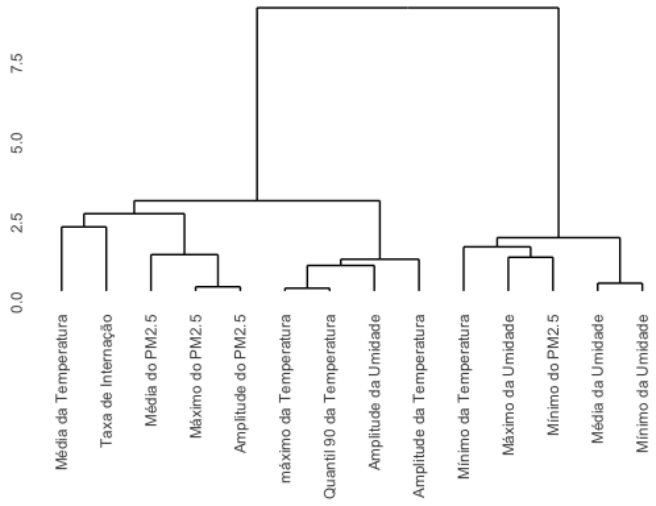


Figura 1: Dendrograma contendo todas as variáveis de pesquisa

Através do dendrograma, que foi utilizando o método de Ward, observamos a proximidade entre a variável Taxa de Internação, média da temperatura e média da concentração de $PM_{2.5}$.

Os mapas da concentração média de $PM_{2.5}$ estão divididos em 3 intervalos, onde o primeiro varia entre 0 e $10\mu g/m^3$, que é o limite de exposição ao material particulado fino para a média anual, o segundo varia entre 10 e $25\mu g/m^3$, onde $25\mu g/m^3$ é o limite diário de exposição ao material particulado fino, e por fim os valores acima de $25\mu g/m^3$. Percebe-se então que nos meses de Agosto e Setembro é onde tem mais municípios com a concentração média de material particulado fino acima do limite diário de exposição.

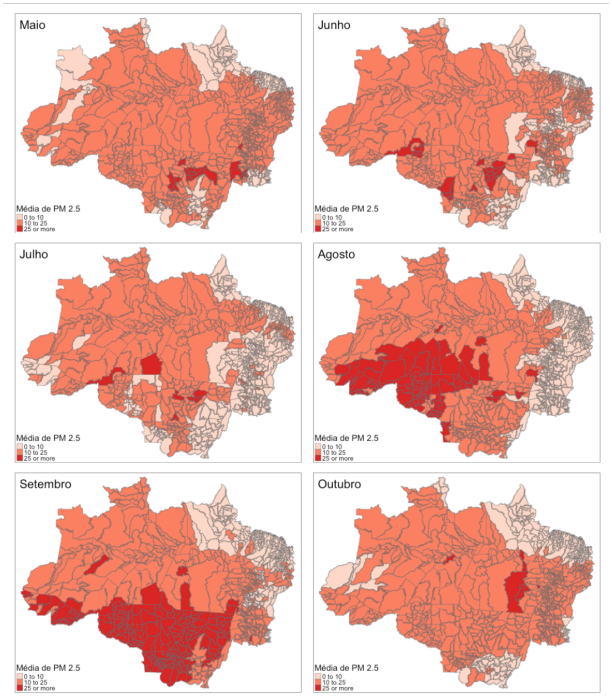


Figura 2: Mapas da concentração média de $PM_{2.5}$ por mês

Os mapas da figura 3, do máximo da temperatura em cada mês também estão divididos em 3 intervalos, onde o primeiro vai de 25 à 30, o segundo de 30 à 35 e o último de 35 à 43. Percebe-se que nos meses de Agosto à Outubro houve uma predominância da temperatura acima de 35 graus Celsius. A variável máxima da temperatura foi escolhida pois a base de dados estava inicialmente com 4 valores por dia, sendo os horários 6h, 12h, 18h e 0h, com isso, o valor médio não havia muita diferença pois equilibrava os valores. Como existe uma grande diferença entre a temperatura observada 12h e às 0h o valor médio não refletia a realidade.

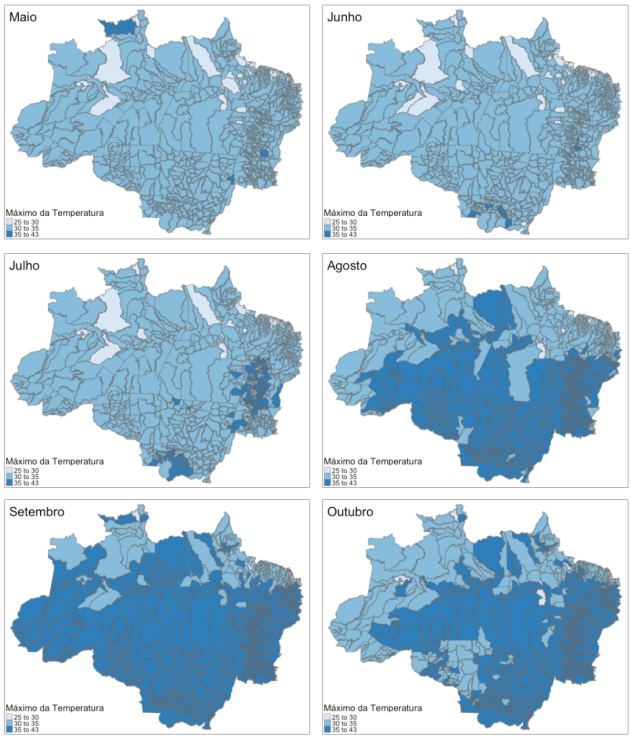


Figura 3: Mapas da temperatura máxima por mês

Sobre a umidade foi selecionada a variável mínimo da umidade para a execução dos mapas. Onde o primeiro intervalo vai de 5 ate 30, onde 5 é menor que o valor mínimo observado e 30 é a porcentagem mínima recomendada pela OMS, o segundo intervalo está entre 30 e 50, e o terceiro entre 50 e 78, sendo que de acordo com a OMS a umidade ideal deve estar em 50 e 80%. Sendo assim, observamos que nos meses de Agosto e Setembro a umidade muito baixa, abaixo de 30%, predominou no território da Amazônia Legal no ano de 2019.

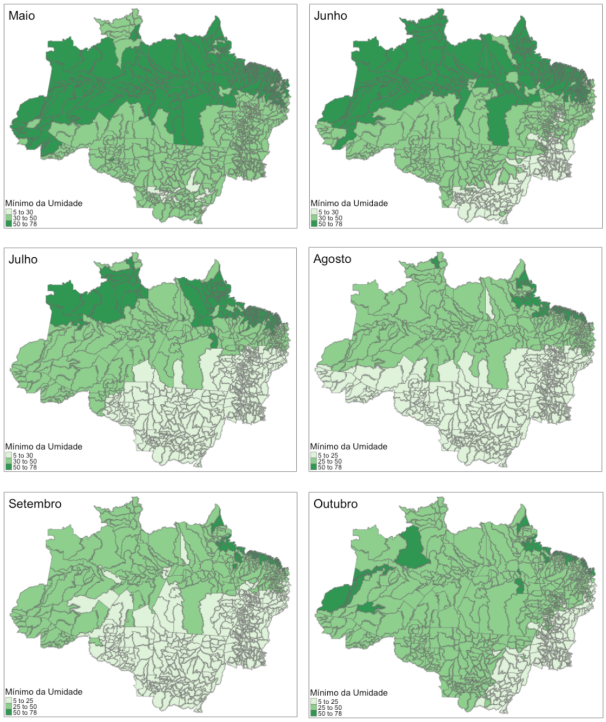


Figura 4: Mapas da umidade mínima por mês

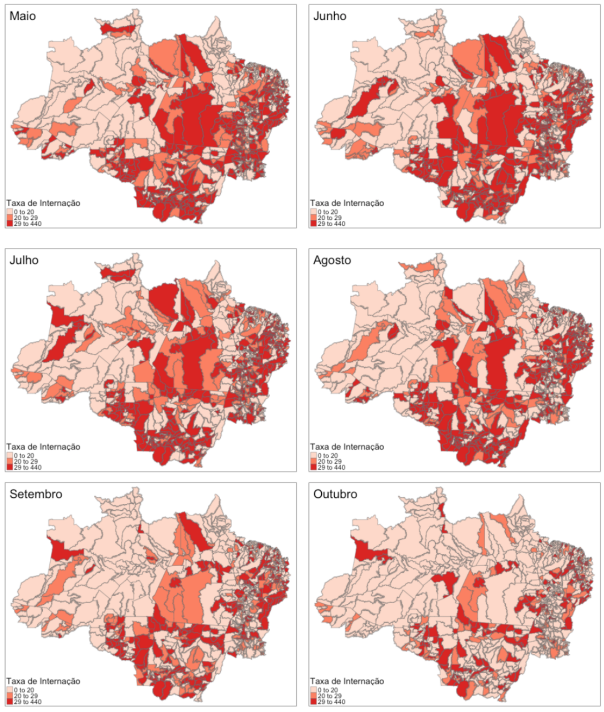


Figura 5: Mapas da Taxa de Internação por mês

Sobre os mapas da Taxa de Internação, que estão apresentados da figura 5, foi dividido em 3 intervalos onde o primeiro vai de 0 à 20, o segundo de 20 à 29 e o terceiro acima de 29, onde 29 é a média da Taxa de Internação da amostra. É observado que nos meses de Maio à Agosto é onde mais municípios tiveram a Taxa de Internação acima da média.

Concluímos então que cada variável varia de acordo com os meses e municípios. Sendo nos meses de Agosto e Setembro onde houve maior concentração média de $PM_{2.5}$ em mais municípios, temperatura máxima com valores mais elevados e maior percentual de municípios com umidade relativa abaixo da recomendada pela OMS. Já a respeito da Taxa de Internação observou-se que nos meses de Maio à Agosto foi onde houve mais municípios com Taxa de Internação acima da média. Este estudo sugere que o mês de Agosto pode ser o mais crítico de risco à saúde em grande parte dos municípios da Amazônia Legal, com centrações altas de $PM_{2.5}$, extremos de calor, baixa umidade relativa e aumento na internação por doenças cardiovasculares.

Referências

[1] da Silva Viana Jacobson, L., de Oliveira, B. F. A., Perez, L. P. e de Souza Hacon, S. Impacto do aquecimento global nos anos potenciais de vida perdidos por doenças cardiorrespiratórias em capitais brasileiras. *Sustentabilidade em Debate 11*, 3 (2020), 304.

[2] IBGE. *Amazônia Legal*, Julho 2021.

[3] INPE. *Focos por Estado*, Outubro 2019.

[4] Johnson, R. A. e Wichern, D. W. *Applied Multivariate Statistical Analysis*, 6 ed. Pearson Education, Inc., 2007.

[5] Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H. e Amann, M. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the global burden of disease study 2010. *Lancet* (2013), 2224–2260.

[6] OPAS. *OMS revela principais causas de morte e incapacidade em todo o mundo entre 2000 e 2019*, Dezembro 2020.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

Análise de roubos na cidade do Rio de Janeiro via modelos aditivos generalizados

Julia Ferreira (UFF)
Aline Pereira (UFF)
Dayana Gimenes (UFF)
Beatriz Pinna (ISP-RJ)
Rafael Erbisti (UFF)
Jony Arrais Pinto Junior (UFF)

Email de contato: juliaferreira@id.uff.br, alinedavila@id.uff.br, dayanagimenes@id.uff.br, beatrizpinna.isp@gmail.com, rerbisti@id.uff.br, jarrais@id.uff.br

Resumo

Com a rápida evolução da pandemia ocasionada pelo novo coronavírus (SARS-CoV-2), medidas foram criadas para diminuir o contágio e a disseminação da doença, dentre elas o isolamento social. Essa restrição à circulação da população promoveu alguns efeito, em especial, na violência das grandes cidades. Nesse sentido, este trabalho busca avaliar e comparar o comportamento da dinâmica espacial dos roubos no município do Rio de Janeiro em abril e maio de 2019 e 2020. A partir da estimação de modelos aditivos generalizados (GAM), observou-se pouca mudança nas regiões de ocorrência, uma vez que as maiores probabilidade de roubos encontram-se geralmente em regiões em que há grandes vias, como Avenida Brasil e Linha Vermelha. Entretanto, foi observada mudança substancial na magnitude das probabilidades de roubo entre os meses de abril de maio de 2019 e 2020.

Palavras-chave: modelos aditivos generalizados, espacial, criminalidade, roubos, COVID-19.

Introdução

Em novembro de 2019, na província chinesa de Wuhan, foi detectado pela primeira vez o novo coronavírus (SARS-CoV-2). Em pouco tempo, o vírus se espalhou por todo o mundo e em 11 de março de 2020 a Organização Mundial da Saúde caracterizou a COVID-19 como uma pandemia global. No Brasil, o primeiro caso de COVID-19 foi confirmado no Estado de São Paulo no dia 26 de fevereiro de 2020. O primeiro óbito ocorreu em 17 de março, também em São Paulo [2]. Com o objetivo de evitar um colapso sanitário nos países, devido à rápida evolução da pandemia, medidas foram criadas para diminuir o contágio e a disseminação da doença, dentre elas o isolamento social [6].

Tanto no estado quanto na cidade do Rio de Janeiro foram impostas medidas para reduzir a movimentação e aglomeração de pessoas nas ruas. Os decretos estadual e municipal que discutiam e proibiam a circulação da população, impedindo o uso de transportes públicos, aulas presenciais, eventos e abertura de estabelecimentos que prestavam serviços não essenciais iniciaram em 17 de março de 2020 e permaneceram inalterados até final de maio de 2020[1].

Segundo [6], o isolamento social é uma estratégia não farmacológica que incluiu o fechamento de escolas e universidades, comércio não essencial e áreas públicas de lazer. Nessa perspectiva, esse processo de isolamento social promoveu alguns impactos, em especial, na violência. Um estudo realizado em 2020, nas cidades de Los Angeles e Indianópolis, evidenciou uma diminuição dos crimes devido à quarentena imposta pela COVID-19, principalmente nas ocorrências de roubos [5].

De forma geral, a criminalidade é sempre preocupante tanto para a população que esta vulnerável como para os gestores, pois afeta diretamente o bem-estar dos indivíduo trazendo perdas monetárias e pessoais. Nesse contexto, a criminalidade gera custos para a sociedade. De acordo

[1] A partir de 1º de junho de 2020, tanto a prefeitura do Rio de Janeiro quanto o estado do Rio iniciaram a flexibilização das medidas restritivas.

com o Atlas da Violência de 2019 [1] a criminalidade gerou uma perda para o Brasil, em 2016, de 5,9% do PIB. Esse custo do crime afeta diretamente a entrada de novos investimentos e provoca perdas nas atividades turísticas. No caso específico do Rio de Janeiro, os problemas com segurança pública na cidade assombram a população há muito tempo.

Nesse sentido, o objetivo deste trabalho é avaliar e comparar o comportamento da dinâmica espacial dos tipos de roubos mais comuns no município do Rio de Janeiro em abril e maio de 2019 e 2020. Busca-se então verificar se o isolamento social causado pela COVID-19 mudou a dinâmica espacial ou diminuiu a probabilidade de ocorrência desses roubos na cidade.

Materiais e Métodos

O Instituto de Segurança Pública do Estado do Rio de Janeiro (ISP-RJ) é a instituição responsável pelo monitoramento, análise e divulgação de estudos sobre a criminalidade no Estado do Rio de Janeiro. Os dados utilizados neste trabalho foram disponibilizados pelo ISP-RJ[2].

Os dados contemplavam diversos tipos de ocorrências policiais no período de 2016 a 2020. Aqui, nos concentraremos apenas nos meses de abril e maio dos anos de 2019 e 2020. O recorte temporal feito baseia-se no interesse de comparar a ocorrência de determinado crime no período de início de isolamento social causado pela pandemia do novo coronavírus com o mesmo período do ano anterior. Sendo assim, de forma específica, nosso interesse é avaliar a ocorrência de tipos de roubos mais comuns[3] na cidade do Rio de Janeiro, comparando os períodos de abril e maio de 2019 e 2020.

No contexto espacial, os dados dos tipos roubos mais comuns na cidade do Rio de Janeiro foram disponibilizados por quadrículas de 200×200 metros, ou seja, há informação sobre a quantidade de roubos ocorridos em cada quadrícula que pertence ao município. A cidade do Rio de Janeiro possui 30.451 quadrículas. A Figura 1 apresenta um recorte do município do Rio para ilustrar os bairros de Copacabana e Leme divididos em quadrículas. Devido à grande dimensão das informações,

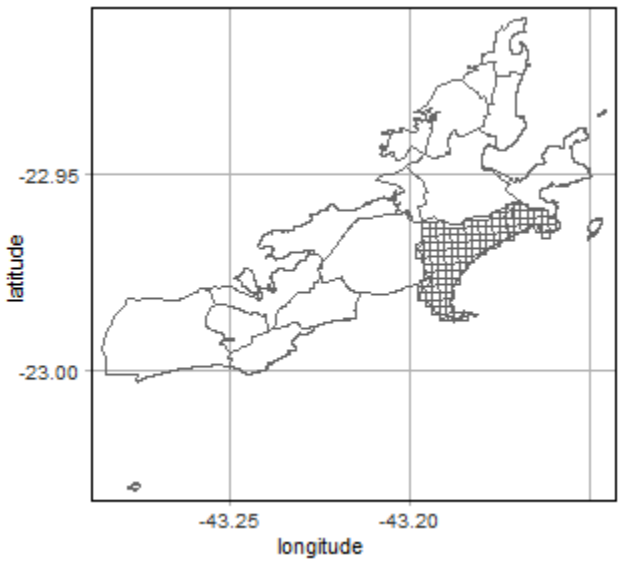


Figura 1: Quadrículas de 200×200 metros dos bairros Copacabana e Leme, Zona Sul da cidade do Rio de Janeiro.

utilizar qualquer modelo estatístico para capturar a dependência espacial dos dados a partir de componentes estruturadas torna-se inviável. Nesse sentido, podemos buscar outros métodos que considerem de alguma forma a associação espacial que há nas observações.

Os Modelos Aditivos Generalizados (GAM) [3, 4] podem ser utilizados na modelagem espacial, pois adicionam funções de suavização aos preditores fornecendo grande flexibilidade. Desta forma, seja Y_{it} a ocorrência de roubo na quadrícula i no tempo t, $i = 1, \ldots, 30.451$ e $\{(s_{1i}, s_{2i}) : i =$

[2] A obtenção dos dados foi feita via Convênio de Cooperação Técnica Nº 9734794/2020, firmado entre a Universidade Federal Fluminense e o Instituto de Segurança Pública do Estado do Rio de Janeiro.

[3] Os tipos de roubos mais comuns foram categorizados a partir de três tipos de ocorrências descritas nas bases de dados disponibilizadas pelo ISP/RJ: Patrimônio-violento-móvel, Patrimônio-violento-rua e Patrimônio-violento-fixo.

$1, \ldots, 30.451\}$ o conjunto de coordenadas, onde s_{1i} é a longitude e s_{2i} a latitude do centroide da quadrícula i, e t é o período de tempo avaliado, com $t = 1$ indicando o mês de abril de 2019, $t = 2$, maio de 2019, $t = 3$, abril de 2020 e $t = 4$, maio de 2020.

Sendo assim, considere o GAM para a ocorrência do evento de interesse, ocorrência de alguns tipos de roubos, $Y_{it} = 1$, em cada momento do tempo avaliado t:

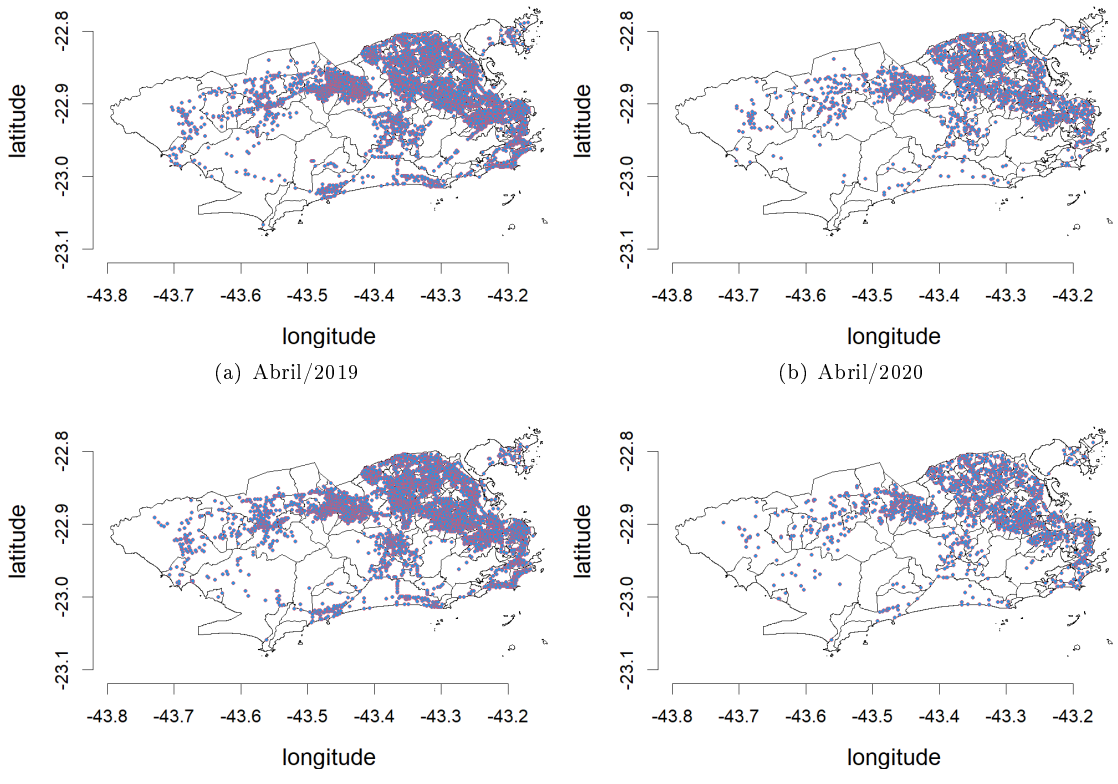
$$Y_{it} \sim \text{Bernoulli}(\pi_{it}), \quad i = 1, \ldots, 30.451; \quad t = 1, \ldots, 4$$
$$ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta + \psi(lat_i, long_i), \tag{1}$$

onde π_{it} é a probabilidade de ocorrência de roubo no centroide i no período de tempo t, β é o nível global, ψ é uma função que captura relações não lineares, mais especificamente, é uma função na base da *spline* cúbica, e $\psi(lat_i, long_i)$ é o componente relativo à posição geográfica do centroide da célula, atuando como suavizador de um efeito espacial.

A estimação das quantidades desconhecidas do modelo é feita a partir da maximização de verossimilhança penalizada. Todo o procedimento de ajuste dos modelos foi feito no *software* `R`, a partir das funções do pacote `mgcv`.

Resultados e Discussão

O período de análise compreende os meses de abril e maio dos anos de 2019 e 2020. A Figura 2 apresenta as localizações em que houve ocorrência de roubo para os períodos avaliados. Nota-se que os pontos de ocorrência são similares em ambos os meses tanto em 2019 quanto em 2020. Entretanto, é evidente o menor número de localizações com ocorrências de roubos em 2020. Outro ponto importante é a maior concentração de roubos nas áreas onde há as principais vias expressas da cidade, como a Linha Amarela, Linha Vermelha e Avenida Brasil. Em abril e maio de 2019



(a) Abril/2019 (b) Abril/2020

(c) Maio/2019 (d) Maio/2020

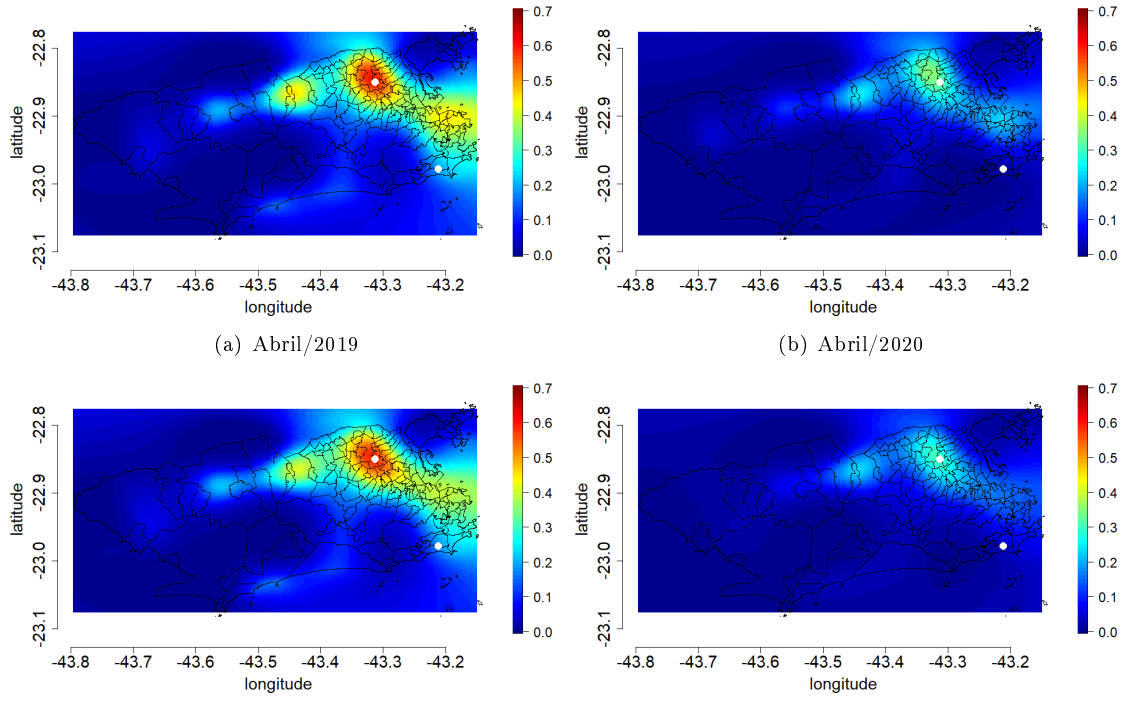
Figura 2: Localizações com ocorrência dos tipos de roubos mais comuns na cidade do Rio de Janeiro.

houve ocorrência de roubos em 22,0% e 21,7% das quadrículas avaliadas, respectivamente. Para os

mesmos meses de 2020, esses números foram 7,4% e 6,8%. Essa diminuição considerável também pode ser vista na Figura 2.

O modelo descrito na equação (1) foi ajustado para cada mês avaliado. A Figura 3 apresenta a probabilidade estimada de ocorrência dos tipos de roubos mais comuns na cidade do Rio de Janeiro. Observe que não parece haver mudança nas regiões de ocorrência. De forma geral, as maiores probabilidades dessa seleção de roubos estão nas regiões em que há grandes vias, como Avenida Brasil e Linha Vermelha. Entretanto, note que a magnitude das probabilidades é bastante diferente entre os meses de 2019 e 2020.

A Figura 3 também apresenta dois pontos: um localizado em Vicente de Carvalho, bairro com altos índices de roubo (ponto branco localizado ao norte da imagem), e outro na Lagoa, bairro nobre do Rio de Janeiro (ponto branco localizado ao sul da imagem)[4]. Avaliando a razão de chances entre essas duas localizações na cidade, temos que: em abril de 2019, o risco de roubo em Vicente de Carvalho era 6,3 vezes maior do que o risco de roubo na Lagoa. Em maio de 2019, esse risco foi de 6,8. Quando avaliamos os meses durante o período de isolamento, o risco de roubo em Vicente de Carvalho foi de 31,3 e 10,1 vezes maior do que o risco de roubo na Lagoa para os meses de abril e maio de 2020, respectivamente. Note que o risco em abril é substancialmente elevado, uma vez que os tipos de roubos estavam bastante concentrados naquela região do subúrbio do Rio. Em maio de 2020, apesar do quantitativo dos tipos de roubos ser menor do que em abril, houve maior dispersão das ocorrências pelo território da cidade. Os resultados apresentados aqui podem ser facilmente estendidos para outras regiões da cidade. Além disso, nota-se importante relação temporal a ser considerada numa modelagem mais completa. Espera-se que as análises feitas contribuam para a discussão de segurança pública na cidade do Rio de Janeiro, podendo gerar políticas públicas eficazes ao combate da criminalidade na cidade.



(a) Abril/2019 (b) Abril/2020

(c) Maio/2019 (d) Maio/2020

Figura 3: Probabilidade de ocorrência dos tipos de roubos mais comuns na cidade do Rio de Janeiro estimada. Pontos brancos: localização no bairro de Vicente de Carvalho (ponto ao norte da imagem) e localização no bairro da Lagoa (ponto ao sul da imagem).

Referências

[1] Cerqueira, D., Bueno, S., Lima, R. S., Neme, C., Ferreira, H., Alves, P. P., Mar-

[4]Os bairros foram identificados de forma aproximada a partir dos pontos ilustrados nos mapas

ques, D., Reis, M., Cypriano, O., Sobral, I., Pacheco, D., Lins, G. e Armstrong, K. *Atlas da violência 2019*. Brasília: Rio de Janeiro: São Paulo: Instituto de Pesquisa Econômica Aplicada; Fórum Brasileiro de Segurança Pública., 2019.

[2] da Saúde, M. Painel coronavírus. 2021.

[3] Hastie, T. e Tibshirani, R. Generalized additive models (with discussion). *Statistical Science 1* (1986), 297–318.

[4] Hastie, T. e Tibshirani, R. *Generalized Additive Models*. Chapman & Hall, 1990.

[5] Mohler, G., Bertozzi, A. L., Carter, J., Short, M. B., Sledge, D., Tita, G. E., Uchida, C. D. e Brantingham, J. Impact of social distancing during covid-19 pandemic on crime in los angeles and indianapolis. *Journal of Criminal Justice 68* (1979).

[6] Y. Wang, Y. Wang, Y. C. e Qin, Q. Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (covid-19) implicate special control measures. *Journal of Medical Virology 92* (2020), 568–576.

Eventos e ondas de calor e a internação por morbidades cardiovasculares e respiratórias no bairro de Irajá/RJ

Juliana Vilardo Mendes (UFRJ)
Leonardo Caçadini Bizerra da Silva (UFRJ)
Nubia Beray Armond (UFRJ)
Ludmilla da Silva Viana Jacobson (UFF)
Rafael Erbisti (UFF)

Email de contato: julianavilardo@outlook.com, leocacadini@gmail.com, nubia.beray@gmail.com, ludmillajacobson@id.uff.br, rerbisti@id.uff.br

Resumo

Num cenário de mudanças climáticas globais, eventos extremos, como ondas de calor, têm se mostrado cada vez mais frequentes e intensos. O aumento na exposição de determinados grupos sociais a esses eventos têm desencadeado impactos na saúde, principalmente em relação à morbidades do aparelho respiratório e circulatório. O objetivo deste trabalho é analisar a influência de eventos de calor (EC) e ondas de calor (OC) nas internações por doenças do aparelho circulatório e respiratório no bairro de Irajá, Rio de Janeiro, no período de 2015 a 2019. Na análise dados, as ondas foram definidas a partir da duração de três e sete dias consecutivos, com temperaturas superiores aos percentis 90% e 75%. Os dados de internações foram obtidos pelo Sistema de Informações Hospitalares do SUS (SIH/SUS). Para avaliar a associação entre temperatura e saúde, foi utilizado o modelo de defasagem distribuída não linear a partir do pacote `dlnm` do *software* `R`. Os resultados obtidos mostram que limiares mais elevados sugerem risco aumentados de internações, além de caracterizar um perfil da população mais exposta.

Palavras-chave: ondas de calor, modelo de defasagem distribuída não linear, aparelho circulatório.

Introdução

Num cenário de mudanças globais, estudos indicam que eventos extremos, como ondas de calor, estão se tornando cada vez mais intensos, frequentes e duradouros [8, 10]. De acordo com a Organização Mundial da Saúde, a exposição a ondas e eventos de calor deflagra impactos na saúde humana, que variam conforme sua intensidade e duração, além de fatores como a infraestrutura e adaptação. Quando exposto ao calor, o sistema de termorregulação humano reage de modo a compensar a temperatura do meio externo, visando o equilíbrio entre a produção e a perda de calor. A partir de certos limiares, essa relação pode se tornar danosa, uma vez que o corpo humano, na maioria das vezes, responde negativamente, podendo agravar e auxiliar no desencadeamento de morbidades e aumentando o risco de internações e óbitos [11], principalmente por doenças do aparelho respiratório e circulatório [6]. Crianças e idosos, assim como pessoas com doenças crônicas que tomam medicação diariamente, possuem um risco maior de desenvolver complicações e de morrer durante uma onda de calor.

Entre os bairros que possuem estações meteorológicas na Cidade do Rio de Janeiro, o Bairro de Irajá tem se destacado pelas temperaturas elevadas. Nos últimos anos, inclusive, por conta de fatores como a pouca arborização, alto fluxo de veículos e elevada densidade de construções, foi o bairro em que mais foram identificadas ondas de calor [7, 9]. Desse modo, o presente trabalho tem como objetivo analisar a influência de ondas de calor (OC) e eventos (EC) nas internações por doenças do aparelho circulatório e doenças do aparelho respiratório (Capítulos IX e X da Classificação Internacional das Doenças - CID), no bairro de Irajá, dentre a série histórica de 2015 a 2019.

Materiais e Métodos

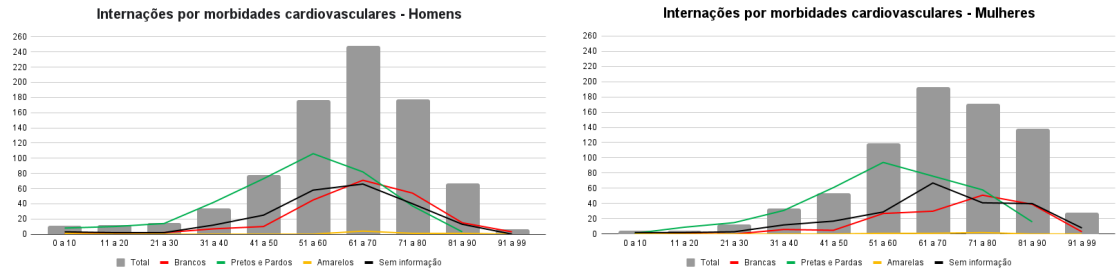
Em relação ao recorte espacial, o bairro de Irajá está situado na Zona Norte do município do Rio de Janeiro, localizado no litoral sudeste brasileiro. Compreende uma área de 7,4 km2, e abrange uma população estimada de 96.382 habitantes, segundo dados da projeção populacional do Instituto Brasileiro de Geografia e Estatística, sendo considerado como bairro de porte médio. Quanto à sua posição geográfica, Irajá faz limite com 16 bairros adjacentes, é atravessado pela Avenida Brasil e está situado na retaguarda do maciço da Tijuca.

Neste trabalho, para a identificação das OC e EC, foi adaptada a metodologia de [12], no qual foram testados dias com temperaturas superiores aos percentis 99%, 98%, 97%, 90% e 75%, para os eventos de calor e para ondas de calor. No entanto, foi definido somente a utilização dos percentis 90% e 75% e usados intervalos de 3 e 7 dias consecutivos, respectivamente, com temperaturas superiores aos percentis, definidos a partir de valores médios mensais. Os percentis foram extraídos dos dados mensais da série histórica, e não dos dados anuais, conforme parte significativa da literatura apresenta. Essa escolha se deu por conta da maior previsibilidade dos meses de verão em apresentar temperaturas mais elevadas em relação aos meses de inverno. Assim, foram utilizados dados horários de temperatura média do ar entre 2015 e 2019 da estação meteorológica do Sistema AlertaRio[1],

Em relação aos dados de internações, foram obtidos pelo Sistema de Informações Hospitalares do Ministério da Saúde (SIH/SUS), e através do TabWin, foram obtidos a data de internação e o CEP de residência dos pacientes, que foram filtrados para a área de estudo. Para avaliar a associação entre temperatura e saúde, foram elaborados gráficos para exemplificar o perfil da população exposta. A partir do modelo de defasagem distribuída não linear (DLNM) [4] foi possível ajustar simultaneamente tanto a tendência quanto a defasagem da exposição por uma estrutura não linear, conhecida como ajuste da associação exposição-defasagem-resposta [4]. Para isso foi utilizado um período de 10 dias para capturar a defasagem nos efeitos do calor sobre as internações. A estimação dos modelos foram feitas a partir do pacote `dlnm` do *software* `R`.

Resultados e Discussão

O bairro de Irajá apresentou a ocorrência de 21 OC e 220 EC (P90), com duração mínima de 3 dias consecutivos, tendo maior concentração de ocorrências de OC e EC, nos meses de Abril e Junho no ano de 2019. Quanto ao P75, foi possível observar a ocorrência de 10 OC e 459 EC, com duração mínima de 7 dias, tendo maior frequência de OC no mês de Abril e EC no mês de Outubro de 2019. Em ambos os limiares, foi identificado que os meses de maiores frequências de eventos e ondas de calor não são característicos do verão, o que justifica a escolha por dados mensais ao invés de anuais. Por serem caracterizados como meses de transição sazonal e apresentarem temperaturas mais estáveis, são mais sensíveis à atuação de sistemas tropicais (mTa) que podem produzir eventos marcados por temperaturas mais elevadas. Os dados de internações demonstram que houveram 2565 internações de residentes do bairro de Irajá durante o período analisado, sendo 1567 referentes a morbidades do aparelho circulatório e 998 morbidades do aparelho respiratório. A partir dos dados de internações por morbidades cardiovasculares apresentados na Figura 1,

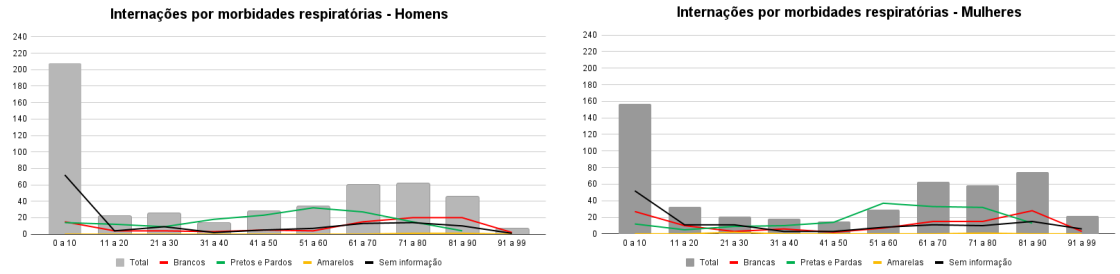


(a) (b)

Figura 1: Número de internações por morbidades cardiovasculares segundo sexo, faixa etária e cor.

[1] http://alertario.rio.rj.gov.br/

percebe-se que, para homens e mulheres, ao envelhecer, aumentam os totais de internação, relação conhecida na literatura [3]. Além disso, a faixa entre os 51-90 anos apresenta-se como a mais vulnerável, especialmente no intervalo 61-80 anos. Uma análise mais minuciosa de faixa etária demonstra que a população masculina de 61-80 anos representa mais da metade (424) do total de internações desse sexo (817). Há um salto significativo de internações entre as faixas 41-50 e 51-60, onde o risco mais do que duplica em relação a faixa anterior, indo de 77 para 176 (homens) e de 53 para 118 (mulheres). Há queda abrupta de internações na faixa de 81-90 entre os homens, o mesmo não é visualizado entre as mulheres por conta de sua maior longevidade. Além disso, no caso das mulheres, a faixa de 61-70 é a única que, comparativamente com a dos homens, tem uma diferença proporcionalmente maior do que a faixa anterior, chegando a passar o número de internações dos 81 anos em diante. Na mulher, o início da doença cardiovascular está associado ao fim da menopausa, sendo tipicamente mais tardio [2]. Atendendo-se às internações por morbidades



(a) (b)

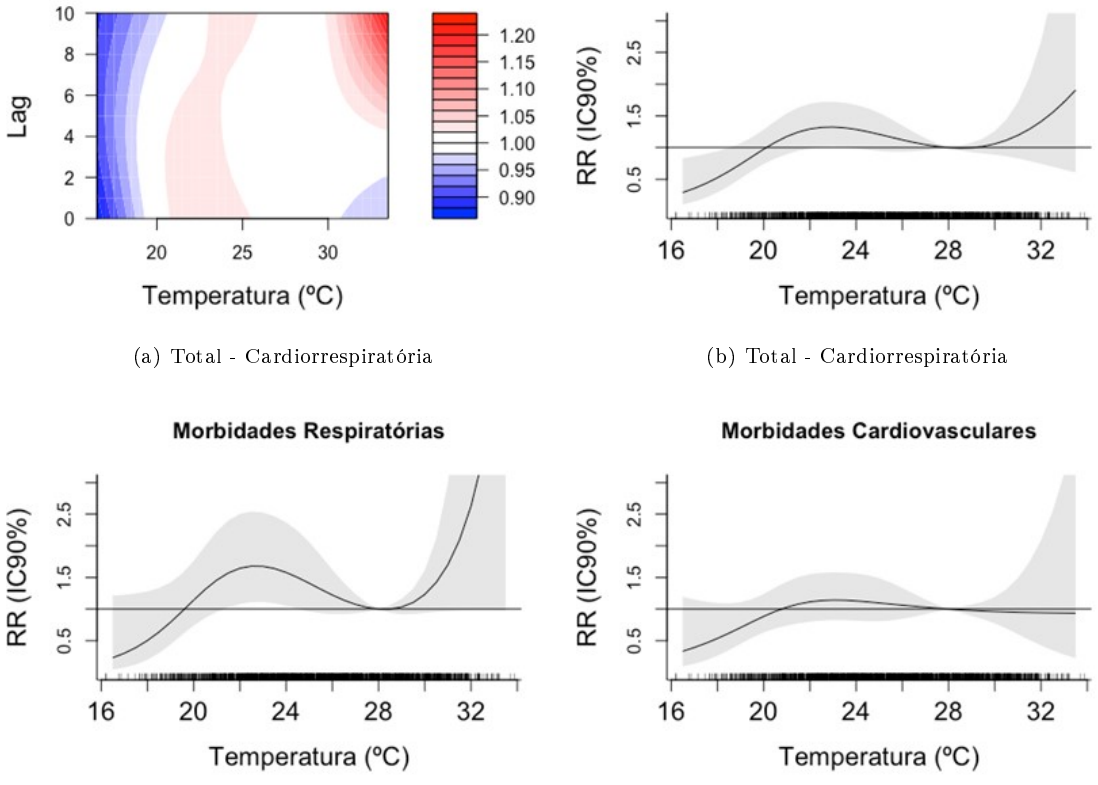
Figura 2: Número de internações por morbidades cardiovasculares segundo sexo, faixa etária e cor.

respiratórias, uma dinâmica diferente é encontrada. A partir da Figura 2 nota-se que a população de 0-10 anos e acima dos 61 anos é a mais afetada. A concentração de internações nessas faixas etárias pode ser explicada pelo fato dessas pessoas serem, além de mais sensíveis a extremos de temperatura, possuírem o estado de saúde fragilizado, ressaltando o aumento da comorbidade em idosos [5] e que crianças até 2 anos de idade não terem o aparelho respiratório completamente desenvolvido [1]. Dentro da faixa de 0-10, inclusive, o total de internações de crianças de até dois anos somam 45,5%. Em ambos grupos de morbidades, percebe-se que há um elevado número de ausência de registro da informação de raça/cor do paciente, sendo 27,6% do total de internações. Além disso, o número das internações foi consideravelmente maior entre a população de pretos e pardos.

Quanto à relação entre a exposição das ondas e seu efeito na ocorrência de internações, foi possível identificar risco aumentado para as temperaturas mais elevadas, acima de 28°C, através do percentil 75%, considerando a defasagem de 10 dias da exposição para a área de estudo, como apresentado na Figura 3. Além disso, observa-se uma curva estimada que sugere efeito colheita. Nas morbidades respiratórias verifica-se também um aumento médio significativo de 58% (IC 90%: 1,05 - 2,37) no risco de internação para a exposição acumulada a temperatura de 24°C, quando comparada a temperatura de 28°C.

A Tabela 1 apresenta os riscos relativos acumulados estimados para o efeito global da onda de calor (representada pela temperatura mediana do período em que houve onda de calor) e o efeito adicional da onda de calor (quando o modelo é ajustado pela temperatura média diária), sobre os desfechos de saúde. Para a maioria dos resultados observa-se um efeito protetor, não significativo. No entanto, os resultados sugerem que as morbidades respiratórias e cardiovasculares têm respostas à exposição diferentes. Nas morbidades respiratórias observa-se um efeito adicional da onda de calor (RR = 1,12) definida como 7 dias de duração e percentil 75 da temperatura. Enquanto que para as morbidades cardiovasculares esse efeito adicional da onda de calor (RR= 1,29) ocorre para a definição de 3 dias de duração e percentil 90 da temperatura.

Outras investigações são necessárias para discutir melhor os resultados deste trabalho, por exemplo, a investigação dos efeitos das ondas de calor separadamente por morbidades respiratória e cardiovascular, assim como por faixas etárias.



(a) Total - Cardiorrespiratória (b) Total - Cardiorrespiratória

(c) Respiratória (d) Cardiovascular

Figura 3: Efeito acumulado em 10 dias da exposição à temperatura em cada morbidade (respiratória e cardiovascular), centrado no percentil 75 da temperatura média diária. RR = risco relativo.

Tabela 1: Efeito acumulado da Onda de Calor na Internação por doenças cardiorrespiratórias, Irajá, Rio de Janeiro, 2015-2019.

Morbidade	Onda de calor		RR	IC 90%
Cardiorrespiratória (cardiovascular + respiratória)	3 dias consecutivos em percentil 90	efeito global da OC (temp. mediana = 29,11ºC)	0,83	(0,61-1,12)
		efeito aditivo da OC	0,99	(0,79-1,25)
	7 dias consecutivos em percentil 75	efeito global da OC (temp. mediana = 28,97ºC)	0,83	(0,62 - 1,12)
		efeito aditivo da OC	0,94	(0,75 - 1,18)
Cardiovascular	3 dias consecutivos em percentil 90	efeito global da OC	0,86	(0,59 - 1,23)
		efeito aditivo da OC	1,29	(0,99 - 1,66)
	7 dias consecutivos em percentil 75	efeito global da OC	0,92	(0,64 - 1,33)
		efeito aditivo da OC	0,83	(0,62 - 1,12)
Respiratória	3 dias consecutivos em percentil 90	efeito global da OC	0,78	(0,49 - 1,24)
		efeito aditivo da OC	0,60	(0,40 - 0,92)
	7 dias consecutivos em percentil 75	efeito global da OC	0,71	(0,45 - 1,13)
		efeito aditivo da OC	1,12	(0,81 - 1,55)

Referências

[1] Aleixo, N. C. Pelas lentes da climatologia e da saúde pública: doenças hídricas e respiratórias na cidade de Ribeirão Preto/SP. Tese de doutorado, Faculdade de Ciências e Tecnologia. Universidade Estadual Paulista, Presidente Prudente/SP, 2012.

[2] Escosteguy, C. C. Epidemiologia das doenças cardiovasculares nas mulheres. *Revista da SOCERJ* (2002).

[3] Filha, T. e Miranda, M. Prevalência de doenças crônicas não transmissíveis e associação com autoavaliação de saúde: Pesquisa nacional de saúde, 2013. *Revista Brasileira de Epidemiologia 18* (2015), 83–96.

[4] Gasparrini, A. Distributed lag linear and non-linear models in r: the package dlnm. *Journal of Statistical Software 43*, 8 (2011), 1–20.

[5] Gomes, L. Fatores de risco e medidas profiláticas nas pneumonias adquiridas na comunidade. *Jornal de Pneumologia 27*, 2 (2001), 97–114.

[6] Kovast, S., Wolf, T. e Menne, B. Heatwave of august 2003 in europe: provisional estimates of the impact on mortality. *Euro Surveill 8*, 11 (2004).

[7] Lucena, A. J. A ilha de calor na região metropolitana do Rio de Janeiro. Tese de doutorado, Universidade Federal do Rio de Janeiro/COPPE, 2012.

[8] Meehl, G. e Teobaldi, C. More intense, more frequent, and longer lasting heat waves in the 21st century. *Science 305* (2004), 994–997.

[9] Mendes, J. V. Ondas de calor e de frio no município do Rio de Janeiro (2015 - 2019). In *XIV Simpósio Brasileiro de Climatologia Geográfica - SBCG, 2021, João Pessoa. Anais do XIV Simpósio Brasileiro de Climatologia Geográfica trabalhos do eixo 1: Climatologia Urbana.* 2021, pp. 122–137.

[10] Robinson, A. Increasing heat and rainfall extremes now far outside the historical climate. *Climate and Atmospheric Science 4*, 1 (2021), 1–4.

[11] Sartori, M. G. B. *Clima e percepção geográfica – Fundamentos Teóricos à Percepção Climática e à Bioclimatologia Humana.* Santa Maria: Editora Palloti, 2014.

[12] Silveira, R. D. Risco climático, vulnerabilidade socioespacial e eventos climáticos extremos relacionados ao calor e ao frio no estado do Rio Grande do Sul – Brasil. Tese de doutorado, Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente/SP, 2013.

Queimadas e a internação por asma nos municípios da Amazônia e Pantanal

Leandro Dias Gomes de Carvalho (UFF)
Ludmilla da Silva Viana Jacobson (UFF)
Sandra de Souza Hacon (ENSP - FIOCRUZ)

Email de contato: ledias1998@gmail.com, ludmillajacobson@id.uff.br, sandrahacon@gmail.com.

Resumo

No Brasil, grande parte das queimadas é realizada pelo homem por diversas razões como o desmatamento, as disputas de terras e os protestos e muitas das vezes empobrecem o solo e destroem a fauna e flora local. Como consequência, podem levar à extinção e morte de diversos animais e plantas nativas, assim como à emissão de poluentes no ar. No âmbito municipal e regional a fumaça das queimadas pode provocar diversos problemas de saúde, como a asma. O objetivo principal deste trabalho estimar um modelo preditivo de internações por asma em crianças e adolescentes (com 14 anos ou menos) assim como para as variáveis ambientais (Focos de queimadas, e Material Particulado Fino) nos biomas Amazônia e Pantanal. O período de análise do estudo se refere a Jan./2010 até Dez./2020. Na análise dos dados, foram estimados Modelos de Séries Temporais para avaliar a tendência das variáveis, a partir do SARIMA. Os resultados mostraram que para o Pantanal o número de Focos de Queimadas observados em 2020 foi muito superior ao esperado pelo comportamento da série temporal. Por outro lado, para ambos os biomas o número observado de internações por asma foi muito inferior ao que seria esperado de acordo com o modelo de séries temporais.

Palavras-chave: Asma, Exposição ambiental, Poluição, Queimadas, Séries temporais.

Introdução

A asma é uma doença inflamatória obstrutiva crônica que se caracteriza através da hiperresponsividade das vias aéreas e pela limitação do fluxo aéreo. Se manifesta a partir de episódios recorrentes de falta de ar, aperto no peito, tosse ou sibilância. As crises asmáticas podem ocorrer em consequências de mudanças bruscas de temperatura, poluição ambiental, infecções respiratórias e outros fatores [10].

As queimadas que ocorrem frequentemente tanto no bioma Amazônia quanto no bioma Pantanal, de forma natural ou por meio de ações humanas, liberam uma enorme quantidade de poluentes. Estes poluentes liberados apresentam efeitos diretos na saúde humana, especialmente no sistema respiratório [8]. Além disso, de acordo com a OMS, recomenda-se o limite de $25\mu g/m^3$ para a média diária (24h) de $PM_{2.5}$, para efeitos na saúde humana.

Com relação ao bioma Amazônia, as queimadas estão inseridas no processo produtivo do local, onde esta prática para uso do solo e coleta de madeira é comum e intensificada todo ano durante a estação da seca [7]. Já no âmbito do bioma Pantanal onde a cobertura vegetal predominante é o cerrado, já ocorrem na atividade pecuária, durante a época da estiagem, praticadas com o objetivo de promover a renovação dos alimentos aos rebanhos [6].

Deste modo, o trabalho teve como objetivo principal estimar um modelo preditivo de internações por asma em crianças e adolescentes (com 14 anos ou menos) assim como para as variáveis ambientais (Focos de queimadas, e Material Particulado Fino) nos biomas Amazônia e Pantanal. Os objetivos específicos foram: "Avaliar a tendência da internação por asma em cada bioma e estimar um modelo preditivo de Séries Temporais"; "Comparar e avaliar a diferença entre os resultados observados da internação por asma com os previstos para o ano de 2020".

Material e métodos

A área de estudo foi composta pelos municípios que abrangem os biomas da Amazônia e Pantanal, no período de 2010 a 2020. Para este trabalho foi realizado um estudo ecológico descritivo e de séries temporais utilizando o software R [9] na versão 4.0.3 e a IDE RStudio na versão 1.4.1 e para a análise dos dados foi adotado um nível de significância de 5%.

Foram utilizados os dados fornecidos pelo Instituto Nacional de Pesquisas Espaciais (INPE) [4] sobre o número de Focos de Queimadas. Os dados sobre as concentrações de Material Particulado Fino ($PM_{2.5}$) mensal foram obtidos do Sistema de Informações Ambientais Integrado a Saúde (SISAM) [12]. E os dados sobre o número de internações hospitalares por asma foram obtidos do Sistema de Informações Hospitalares do SUS (SIH/SUS) por meio do Departamento de Informática do Sistema Único de Saúde (DATASUS) [2]. O período de análise do estudo se refere a jan./2010 até dez./2020 e os biomas em estudo foram Amazônia e Pantanal.

Para os modelos de séries temporais foi visto primeiramente sua estacionariedade. Foi necessário verificar e analisar a tendência das séries, suas sazonalidades, seus gráficos de autocorrelação (ACF ou FAC) e autocorrelação parcial (PACF ou FACP) para verificação dos parâmetros que foram estimados nos modelos de cada série através de seus correlogramas. Além disso, foi necessário realizar testes de estacionariedade através do teste de Dickey-Fuller [3]. Foi utilizado a metodologia de Box-Jenkins [1] para fazer a identificação dos parâmetros dos modelos, checar seus diagnósticos e verificar a adequação do modelo escolhido para analisar se os resíduos se assemelham ao erro aleatório através do teste de Ljung-Box [5] e para avaliar a normalidade destes resíduos através do teste de Shapiro-Wilk [11]. Além disso foi utilizado o Critério de Informação de Akaike (AIC) para escolha do melhor modelo, de modo que quanto menor o valor do AIC melhor o ajuste do modelo. E por fim, foram previstos os valores para o ano de 2020 através de modelos SARIMA(p,d,q)(P,D,Q)[α] para Internações por Asma, Focos de Queimadas e $PM_{2.5}$.

Resultados e discussão

Visto que os biomas Amazônia e Pantanal são diferentes entre si, desde o tamanho do território até a vegetação existente no local, foi observado que para o bioma Pantanal, as internações por asma variaram de 2 a 67 enquanto que para o bioma Amazônia já variaram de 608 a 3.142. As variáveis de internação por asma, focos de queimadas e $PM_{2.5}$ médio indicaram possuir sazonalidades tanto através de boxplots mensais quanto através de séries temporais mensais. Nos dois biomas em estudo, a alta de internações ocorre entre os meses de março a junho e a alta de focos de queimadas e $PM_{2.5}$ médio ocorre de agosto a outubro.

Foi visto primeiramente a série temporal de internações por asma no bioma do Pantanal, onde de acordo com sua decomposição aditiva vista na Figura 1 indicou uma tendência decrescente e uma sazonalidade marcante ao longo dos anos.

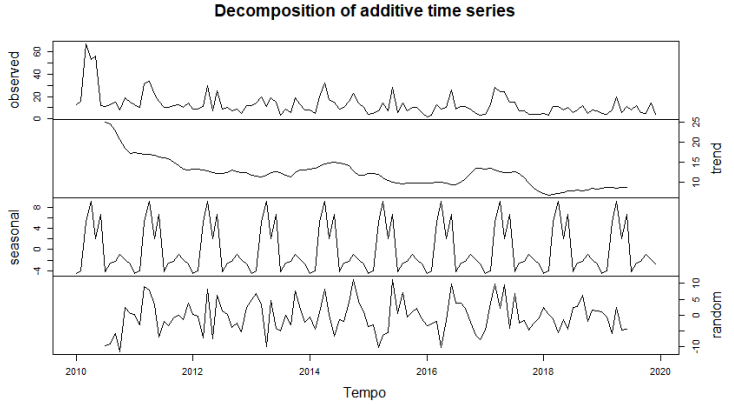


Figura 1: Decomposição aditiva da série temporal de Internações por Asma em crianças e adolescentes de 0 a 14 anos no bioma Pantanal de jan/2010 até dez/2019

O modelo final obtido foi SARIMA(1,1,1)(0,0,1)[12], com AIC de 841,32. O teste de normalidade dos resíduos indicaram que não seguem uma distribuição normal e o teste de Ljung e Box

indicou que não existe autocorrelação nos primeiros 12 lags.

A Figura 2 indica os valores reais da série de internações por asma *versus* os valores ajustados para o período de 2010 a 2019. A linha em preto indica os valores observados no período de 2010 a 2020 e a linha em vermelho indica os valores ajustados do modelo para os anos de 2010 a 2019. A previsão de Internações por Asma em crianças de 0 a 14 anos nos municípios que pertencem ao bioma Pantanal para o ano de 2020 também é dada na Figura 2, onde a sombra cinza escuro indica o intervalo de confiança de 95% para os valores previstos, a linha em preto indica os valores reais analisados nos anos de 2010 a 2020 e a linha em azul indica os valores previstos de Internações por Asma para o ano de 2020.

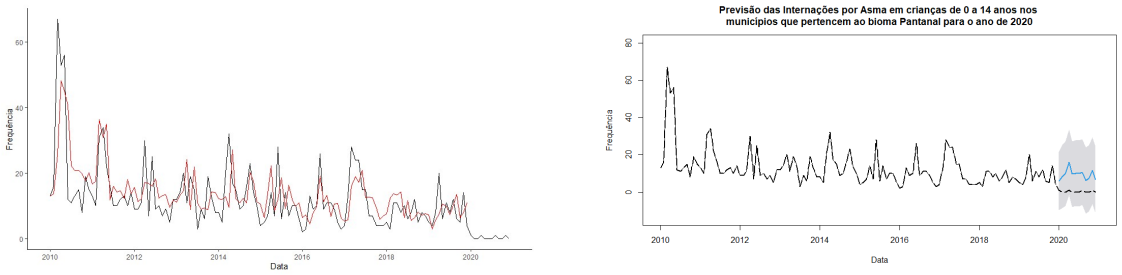


Figura 2: Valores reais *versus* valores ajustados para a série temporal de Internações por Asma em crianças e adolescentes de 0 a 14 anos no bioma Pantanal para o período de 2010 a 2020 e sua previsão para o ano de 2020

Por fim, também foram feitas as previsões para os Focos de Queimadas (SARIMA(1,0,0)(1,1,1)[12]) e $PM_{2.5}$ (SARIMA(0,0,1)(2,0,0)[12]). A Figura 3 indica primeiramente os valores reais *versus* os valores ajustados para o período de 2010 a 2019 e também sua previsão para os Focos de Queimadas para o ano de 2020 no bioma Pantanal e depois os mesmos dois gráficos para o $PM_{2.5}$ médio. Os testes de Shapiro indicaram, nos dois casos, que os resíduos não seguem normalidade e os testes de Ljung e Box também indicaram nos dois casos que não há autocorrelação nos primeiros 12 lags.

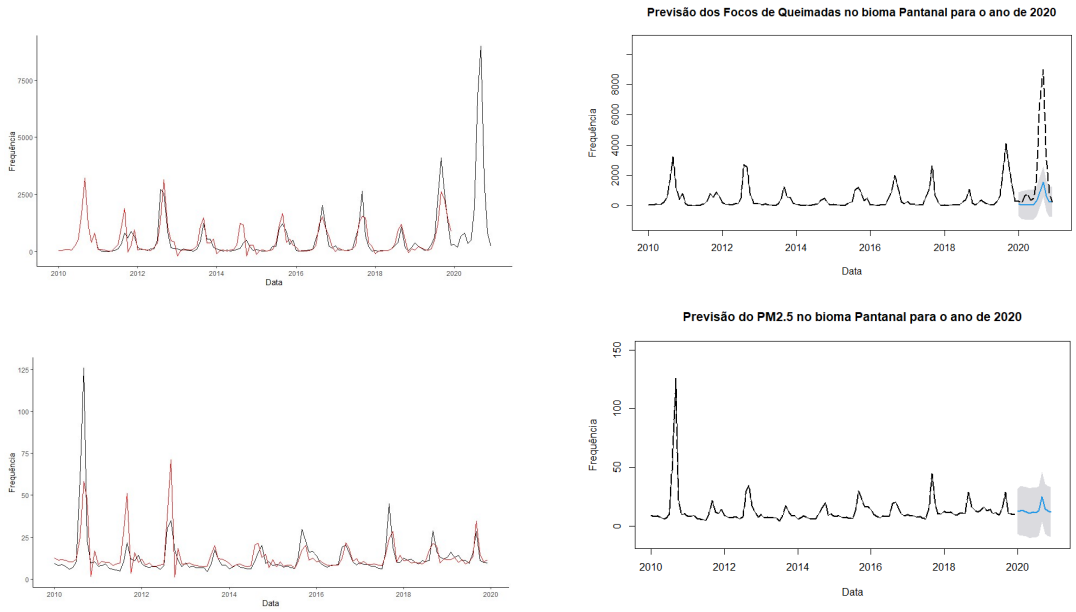


Figura 3: Valores reais *versus* valores ajustados para a série temporal de Focos de Queimadas e $PM_{2.5}$ no bioma Pantanal para o período de 2010 a 2020 e suas previsões para o ano de 2020

Já para o bioma Amazônia, a Figura 4 indica a decomposição aditiva da série temporal de internação por asma, onde também foi observado uma tendência decrescente e uma sazonalidade marcante ao longo dos anos.

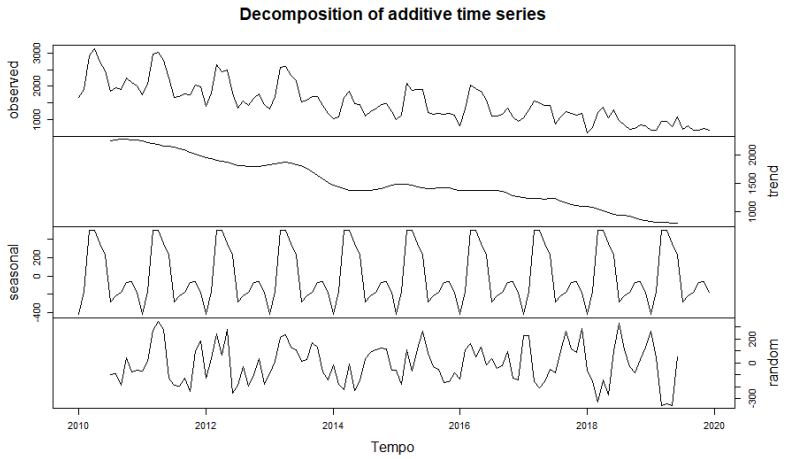


Figura 4: Decomposição aditiva da série temporal de Internações por Asma em crianças e adolescentes de 0 a 14 anos no bioma Amazônia de jan/2010 até dez/2019

O modelo final obtido foi SARIMA(1,1,1)(1,1,1)[12], com AIC de 1423,62. Oo teste de normalidade indicou que os resíduos seguem uma distribuição normal e o teste de Ljung e Box, indicou que não existe autocorrelação nos primeiros 12 lags.

A Figura 5 indica os valores reais da série de internações por asma *versus* os valores ajustados para o período de 2010 a 2019 e a previsão de Internações por Asma em crianças de 0 a 14 anos nos municípios que pertencem ao bioma Amazônia para o ano de 2020.

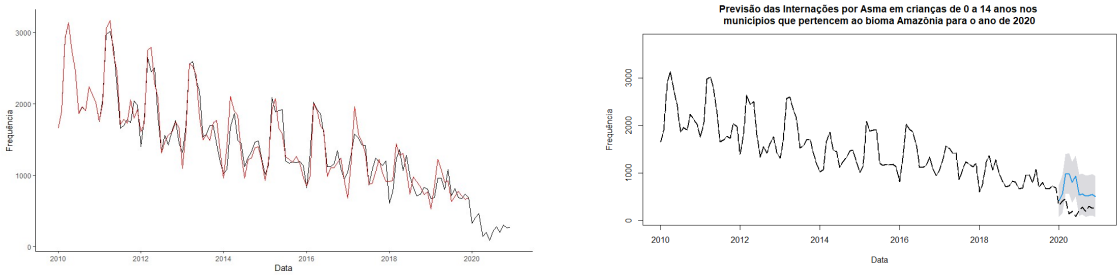


Figura 5: Valores reais *versus* valores ajustados para a série temporal de Internações por Asma em crianças e adolescentes de 0 a 14 anos no bioma Amazônia para o período de 2010 a 2020 e sua previsão para o ano de 2020

Por fim, também foram feitas as previsões para os Focos de Queimadas (SARIMA(1,0,0)(2,1,0)[12]) e $PM_{2.5}$ (SARIMA(1,0,0)(1,1,0)[12]). A Figura 6 indica primeiramente os valores reais *versus* os valores ajustados para o período de 2010 a 2019 e também sua previsão para os Focos de Queimadas para o ano de 2020 no bioma Amazônia e depois os mesmos dois gráficos para o $PM_{2.5}$ médio. Os testes de Shapiro indicaram, nos dois casos, que os resíduos não seguem normalidade e os testes de Ljung e Box também indicaram nos dois casos que não há autocorrelação nos primeiros 12 lags.

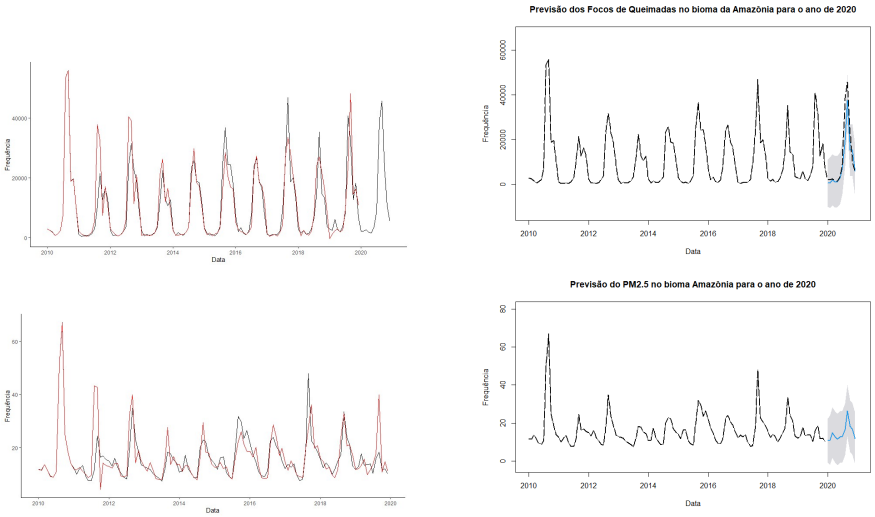


Figura 6: Valores reais *versus* valores ajustados para a série temporal de Focos de Queimadas e $PM_{2.5}$ no bioma Amazônia para o período de 2010 a 2020 e suas previsões para o ano de 2020

Sendo assim, o número de focos verificados em 2020 foi muito superior ao esperado pelo comportamento da série temporal. Já o número observado de internações por asma foi muito inferior ao que seria esperado de acordo com o modelo. Estes resultados podem sugerir que a pandemia de COVID-19 impactou indiretamente na ocorrência e/ou registro de internações por outras causas, assim como nas questões ambientais. Quanto as limitações do estudo, é valido destacar que alguns dos Modelos de Séries Temporais estimados neste trabalho não possuíam normalidade nos resíduos.

Referências

[1] Box, G. E., Jenkins, G. M., Reinsel, G. C. e Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[2] DATASUS. *Internacao por Asma*, 2021.

[3] Dickey, D. A. e Fuller, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association 74*, 366a (1979), 427–431.

[4] INPE. *Inpe Queimadas*, 2021.

[5] Ljung, G. M. e Box, G. E. On a measure of lack of fit in time series models. *Biometrika 65*, 2 (1978), 297–303.

[6] Moraes, E. C., Mataveli, G. A., Santos, P. R. e Oliveira, B. S. Estudo da dinâmica de queimada no bioma pantanal no período de 2002 a 2015. *SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO 18* (2017), 3423–3430.

[7] MOUTINHO, P., Alencar, A., Rattis, L., Arruda, V., Castro, I. e Artaxo, P. Amazônia em chamas: desmatamento e fogo em tempos de covid-19. *Nota Técnica*, 4 (2020).

[8] Nascimento, L. F. C. e Medeiros, A. P. P. d. Internações por pneumonias e queimadas: uma abordagem espacial. *Jornal de Pediatria 88*, 2 (2012), 177–183.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[10] Saldanha, C. e Botelho, C. Queimadas e suas influências em crianças asmáticas menores de cinco anos atendidas em um hospital público. *Rev. bras. alerg. imunopatol 31*, 3 (2008), 108–112.

[11] Shapiro, S. S. e Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika 52*, 3-4 (1965), 591–611.

[12] SISAM. *Sisam Poluicao*, 2021.

Superfície de risco local para casos de dengue, Zika e chikungunya na cidade do Rio de Janeiro

Lucas Moura (UFRJ)
Rafael Erbisti (UFF)
Jony Arrais (UFF)
Nildimar Honório (Fiocruz)

Email de contato: lucasmoura@dme.ufrj.br, rerbisti@id.uff.br, jarrais@id.uff.br, nildimar.honorio@ioc.fiocruz.br

Resumo

Os vírus da dengue, chikungunya e Zika têm registrado crescente incidência e importante expansão geográfica no Brasil, sugerindo limitações para as atuais estratégicas de controle dessas doenças. A dinâmica de transmissão de arboviroses apresenta alta heterogeneidade espaço-temporal, o que torna essencial a identificação de potenciais áreas de maior risco de transmissão. Os modelos estatísticos utilizados usualmente induzem superfície de risco suave, porém, nem sempre essa suposição é razoável. Sendo assim, torna-se relevante buscar estratégias de modelagem que levem em consideração a possível existência de variações locais na autocorrelação espacial. Nesse sentido, o objetivo deste trabalho é avaliar o modelo Binomial-Poisson para o total de casos de arboviroses na cidade do Rio de Janeiro, no período epidêmico de 2019, e incorporar uma estrutura espacial capaz de identificar uma superfície de risco local.

Palavras-chave: modelos CAR, superfície de risco local, arboviroses, inferência bayesiana.

Introdução

Os vírus dengue (DENV), chikungunya (CHIKV) e Zika (ZIKV) são arbovírus causadores de doenças infecciosas emergentes e reemergentes e têm promovido graves epidemias no Brasil e no mundo. O ciclo epidêmico desses arbovírus envolve a transmissão entre humanos por mosquitos do gênero *Aedes*, particularmente *Aedes aegypti* e *Aedes albopictus* (Diptera: Culicidae). A circulação simultânea dos quatro sorotipos do dengue (DENV-1, DENV-2, DENV-3 e DENV-4), CHIKV e ZIKV se tornou uma realidade ao longo dos últimos anos no Brasil, representando um importante desafio para as vigilâncias epidemiológica, entomológica e virológica [8, 12].

Sabe-se que a dinâmica de transmissão dessas arboviroses urbanas depende não apenas do vetor primário, mas de um ambiente permissivo onde vetores, hospedeiros e patógenos interagem [13]. Há diversos estudos que avaliam a dinâmica de transmissão de dengue, chikungunya e Zika, relacionando os dados epidemiológicos com indicadores climáticos, ambientais e entomológicos. Nesse contexto, [2] avaliaram a relação da dengue com indicadores socioeconômicos durante a epidemia de 2001 a 2002 que ocorreu na cidade do Rio de Janeiro, e verificaram a relação entre o percentual de domicílios ligados à rede sanitária e a contagem de casos de dengue. Já [6] discutiram sobre a detecção de clusters persistentes de dengue como uma importante estratégia de vigilância epidemiológica. Na análise feita na cidade do Rio de Janeiro, os autores verificaram que os clusters persistentes de dengue estavam localizados principalmente na Zona Oeste do Rio, região da cidade que possui crescimento urbano desordenado e baixa renda média. Seguindo essa mesma discussão, [14] estratificaram a cidade do Rio de Janeiro em áreas receptivas à dengue, a partir de indicadores que caracterizavam o território e identificaram áreas de maior receptividade à ocorrência de surtos de dengue e alta densidade vetorial. Recentemente, [7] utilizaram modelos espaço-temporais para avaliar a associação de fatores socioambientais e climáticos com casos de chikungunya na cidade do Rio de Janeiro. Ampliando a discussão sobre a dinâmica espacial e temporal de arboviroses, os autores estimaram o efeito da temperatura mínima, a partir de funções de transferência, para verificar o impulso instantâneo dessa variável e sua propagação em tempos futuros nos casos de chikungunya.

De maneira geral, os modelos usualmente utilizados que incorporam efeitos aleatórios espacialmente estruturados, como os utilizados em [15], [5] e [7], induzem superfície de risco suave. Entretanto, nem sempre essa suposição é razoável. Algumas vezes, a superfície de risco estimada pode não ser homogênea no espaço, fazendo com que áreas vizinhas apresentem riscos distintos, especialmente quando estamos avaliando casos de arboviroses numa cidade como o Rio de Janeiro, que apresenta regiões próximas com características sociossanitárias e socioeconômicas distintas. Alguns estudos têm discutido alternativas para esse problema de risco suave, permitindo que a superfície estimada apresente riscos diferentes para regiões vizinhas. Essa discussão é apresentada em [10], [3] [1]. As estruturas discutidas nesses trabalhos são capazes de identificar limites de descontinuidade da superfície de risco e detectar clusters de áreas com comportamentos semelhantes.

Nesse sentido, o objetivo deste trabalho é avaliar o modelo Binomial-Poisson para os casos de dengue, Zika e Chikungunya na cidade do Rio de Janeiro, em 2019, e incorporar estrutura espacial capaz de identificar descontinuidades na superfície de risco estimada.

Materiais e Métodos

Os dados de casos notificados de dengue, Zika e chikungunya foram obtidos a partir do Sistema de Informação de Agravos de Notificação (SINAN) do Ministério da Saúde. As contagens dos casos foram agregadas por bairro ao longo de um período considerado epidêmico, que varia da 9ª até a 33ª semana epidemiológica no ano de 2019, que foi um ano com surto de casos de chikungunya.

Para modelar as variáveis de interesse foram utilizados 5 indicadores: percentual de domicílios ligados a rede de água (IND1), percentual de domicílios adequados (IND2), percentual de domicílios com lixo adequado (IND3), a proporção de cobertura natural (IND4) e a temperatura noturna média do solo em cada bairro durante o período epidêmico (IND5). O três primeiros indicadores foram obtidos a partir dos dados do censo demográfico de 2010, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). A proporção de cobertura natural foi disponibilizada pela Secretaria Municipal de Meio Ambiente do município do Rio de Janeiro. A temperatura noturna do solo foi obtida a partir de imagens de satélites obtidas pela plataforma Google Earth Engine.

Para a construção do método utilizado, considere T_k sendo a contagem do total de casos de arboviroses (dengue, Zika e chikungunya) no k-ésimo bairro no município do Rio de Janeiro. Agora, seja Y_k o número de casos de chikungunya no k-ésimo bairro. Para modelar ambas as variáveis, $\boldsymbol{T} = (T_1, \ldots, T_K)^T$ o total de casos de arboviroses e $\boldsymbol{Y} = (Y_1, \ldots, Y_K)^T$ a contagem de casos de chikungunya, considere o modelo conjunto:

$$p(T_k, Y_k) = p(T_k)p(Y_k \mid T_k), k = 1, \ldots, K. \tag{1}$$

Neste contexto, usualmente, assume-se que o total de casos no bairro k segue uma distribuição Poisson, com uma média definida por E_k e R_k que são o número de casos esperado e o risco de contrair uma das três doenças no k-ésimo bairro, respectivamente, com

$$E_k = \frac{\sum_{k=1}^{K} T_k}{\sum_{k=1}^{K} \text{população}_k} \times \text{população}_k, \quad k = 1, \ldots, K. \tag{2}$$

Além disso, condicionado ao total de casos de arboviroses T_k, assume-se que a distribuição de Y_k é uma binomial com probabilidade de sucesso P_k. Assim, segue que:

$$T_k | E_k, R_k \sim Poisson(E_k R_k), k = 1, \ldots, K \tag{3}$$

$$Y_k | T_k \sim Binomial(T_k, P_k), k = 1, \ldots, K. \tag{4}$$

Para ajustar esse modelo conjunto são atribuídas duas equações para os parâmetros de interesse das distribuições, uma para o risco de se contrair uma das arboviroses e outra para a probabilidade da doença ter sido a chikungunya. De forma geral, essas equações levam em consideração possíveis variáveis explicativas que estão associadas a, pelo menos, um dos parâmetros de interesse, além de um conjunto de efeitos estruturados espacialmente e um conjunto de efeitos aleatórios divididos no espaço. A seguir, apresentamos as formas gerais dessas equações:

$$\log(R_k) = \boldsymbol{x}_{1k}^T \boldsymbol{\beta}_1 + \phi_k^{(1)} + u_k, \tag{5}$$

$$\log\left(\frac{P_k}{1 - P_k}\right) = \boldsymbol{x}_{2k}^T \boldsymbol{\beta}_2 + \phi_k^{(2)} + v_k. \tag{6}$$

A primeira equação é um modelo Poisson para ajustar o risco do total e a segunda equação é um modelo logístico que avalia a probabilidade da ocorrência de chikungunya. $\boldsymbol{x}_{1k}$ e $\boldsymbol{x}_{2k}$ são duas matrizes de variáveis explicativas, um para cada equação. É importante ressaltar que não é necessário utilizar as mesmas covariáveis nessas equações, $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta_2}$ são os coeficientes de regressão associados às variáveis explicativas do risco das doenças e da probabilidade de ser chikungunya, respectivamente. Além disso, $\boldsymbol{\phi}^{(1)} = \left(\phi_1^{(1)}, \ldots, \phi_K^{(1)}\right)^T$ e $\boldsymbol{\phi}^{(2)} = \left(\phi_1^{(2)}, \ldots, \phi_K^{(2)}\right)^T$ são os efeitos estruturados espacialmente que descrevem a autocorrelação não explicada pelas covariáveis, $\boldsymbol{u} = (u_1, \ldots, u_K)^T$ e $\boldsymbol{v} = (v_1, \ldots, v_K)^T$ são os efeitos aleatórios não estruturados. Como as três arboviroses são transmitidas pelo mesmo vetor, é provável que os efeitos espaciais apresentem semelhanças, além disso as duas equações do modelo são ajustadas simultaneamente. Então como em [9], foram considerados que os efeitos espaciais nos dois cenários são proporcionais, isto é, $\boldsymbol{\phi}^{(2)} = a\boldsymbol{\phi}^{(1)}$, em que a é uma constante de proporcionalidade.

Ao seguir uma estrutura Bayesiana, é necessário atribuir distribuições *a priori* para os parâmetros do modelo. Em particular, para os efeitos espaciais $\boldsymbol{\phi}^{(1)}$ é muito comum utilizar distribuições *a priori* da classe CAR (condicionais autorregressivas). Existem diversas *prioris* dessa classe como as propostas em [4] e [11]. Entretanto, além do interesse em modelar as variáveis $\boldsymbol{T}$ e $\boldsymbol{Y}$, gostaríamos de ajustar um modelo capaz de detectar limites de descontinuidades na superfície de risco. Existem casos onde, embora duas regiões sejam vizinhas (contíguas), as duas populações não apresentam uma relação esperada entre vizinhos, em outras palavras, as duas populações são heterogêneas. [10] propuseram um modelo capaz de fazer essa detecção de descontinuidades adaptando a distribuição *a priori* de [11], da seguinte forma:

$$\phi_k^{(1)} \mid \phi_{-k}^{(1)}, \tau, \boldsymbol{\alpha} \sim N\left(\frac{0.99\sum_{j=1}^{K} w_{kj(\boldsymbol{\alpha})}\phi_j^{(1)}}{0.99\sum_{j=1}^{K} w_{kj(\boldsymbol{\alpha})} + 0.01}, \frac{\tau^{-1}}{0.99\sum_{j=1}^{K} w_{kj(\boldsymbol{\alpha})} + 0.01}\right) \tag{7}$$

Observe que esta *priori* considera um parâmetro que mede a correlação espacial global dos efeitos espaciais, ρ, fixo igual a 0.99. Assim, a estrutura espacial pode ser determinada localmente por $w_{jk}(\alpha)$ ao invés de globalmente pelo ρ.

Acredita-se que limites de descontinuidades devem ocorrer entre bairros com populações heterogêneas. Para mensurar isso de alguma forma, são utilizadas as covariáveis do modelo para calcular q medidas de dissimilaridade não negativas $z_{kj} = (z_{kj1}, \ldots, z_{kjq})$, em que $z_{kji} = |z_{ki} - z_{ji}|/\sigma_i$, para $i = 1, \ldots, q$. Cada z_{kji} mede a diferença absoluta no valor de uma covariável entre dois bairros vizinhos k e j. Então, assim como em [10] define-se

$$w_{kj}(\boldsymbol{\alpha}) = \begin{cases} 1 & \text{, se } \exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right) \geq 0.5 \text{ e } j \text{ e } k \text{ são vizinhos;} \\ 0 & \text{, caso contrário.} \end{cases}$$

Nesse caso, em bairros não são vizinhos o $w_{kj}(\boldsymbol{\alpha})$ é zero. Se dois bairros são vizinhos pelo critério de contiguidade, o modelo detectará um risco de limite de descontinuidade se $\exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right) < 0.5$. Os parâmetros de regressão $\boldsymbol{\alpha}$ são restritos a valores não negativos, portanto, quanto maior as medidas de dissimilaridades entre dois bairros maior é probabilidade de que exista um limite de descontinuidade entre eles. Os parâmetros $\boldsymbol{\alpha}$ também possuem limite superior, que para determinado α_i é dado por: $U_i = -\log(0.5)/z_i^{med}$, em que z_i^{med} é a mediana da i-ésima medida de dissimilaridade. Logo, considere $\alpha_i^{min} = -\log(0.5)/z_i^{max}$, em que z_i^{max} é o maior valor da i-ésima medida de dissimilaridade. De acordo com [10], se o limite inferior do intervalo de credibilidade de α_i for maior que α_i^{min}, então a covariável associada a i-ésima medida de dissimilaridade contribui bastante para detectar limites de descontinuidade. Os hiperparâmetros de precisão τ e τ_{uv} são modelados por uma $G(0.001, 0.001)$, enquanto que as *prioris* de μ e a são dadas por uma $N(0, 10)$.

Resultados e Discussão

O conjunto de dados contém aproximadamente 54000 casos de arboviroses notificados da 9ª a 33ª semana epidemiológica no município do Rio de Janeiro em 2019, sendo aproximadamente 34000 casos de chikungunya, 19000 de dengue e 1000 de zika. Campo Grande foi o bairro com o maior número de casos notificados 4000, sendo 2500 apenas de chikungunya. Outros bairros como Bangu e Realengo também apresentaram altos índices das doenças, com mais de 3000 casos cada. Em contrapartida, houve bairros com nenhum caso notificado, como Gericinó, Joá, Parque Columbia e Vasco da Gama.

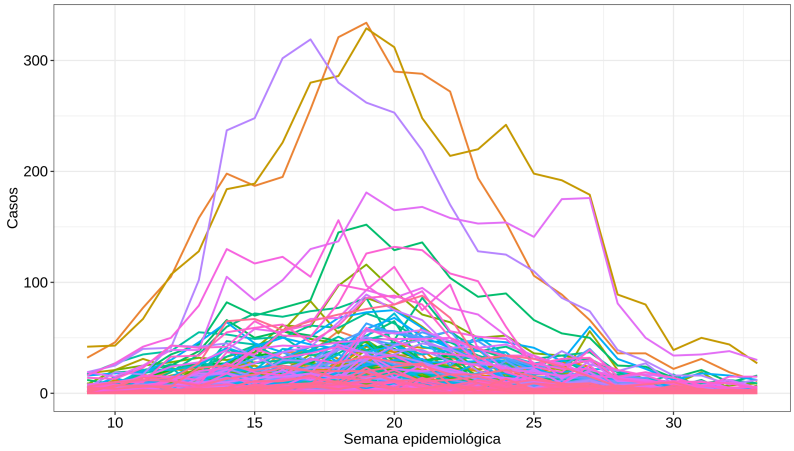


Figura 1: Evolução das notificações de casos de arboviroses em cada bairro da 9ª a 33ª semana epidemiológica.

A Figura 1 apresenta a evolução dos casos de arboviroses nos bairros do Rio de Janeiro.

Note que a partir da 9ª semana epidemiológica, as notificações dos casos crescem gradativamente até atingir o pico de casos por volta da semana 19 e em seguida essas notificações passam a diminuir até a semana 33. Destaca-se ainda que três bairros atingiram um pico superior a 300 casos em uma semana, Realengo na semana 17 e Bangu e Campo Grande na semana 19. A Tabela 1 apresenta algumas medidas descritivas das covariáveis utilizadas no modelo. Pela Tabela 1 observa-se que,

	Min.	Média	Max	DP
IND1	9.09	96.99	100.00	9.837
IND2	4.13	76.11	100.00	21.507
IND3	89.17	99.24	100.00	1.563
IND4	0.00	0.17	0.90	0.187
IND5	18.871	20.28	21.24	0.504

Tabela 1: Medidas resumo como média, desvio padrão (DP), mínimo e máximo das covariáveis usadas para identificar o risco de descontinuidade espacial.

em média 97% dos domicílios estão ligados a rede de água, no entanto, o bairro de Grumari tem um valor de apenas 9% neste indicador, mostrando que as boas condições de saneamento não é a mesma em todos os bairros, como seria o ideal. Além disso, em média 76% dos domicílios são adequados, porém, muitos bairros que são formados por comunidades apresentam valores baixos para este indicador como Mangueira, Complexo do Alemão, Rocinha, e Manguinhos. Por outro lado, todos os bairros apresentam um percentual de lixo adequado acima de 89%. A proporção de área com cobertura natural média é 0.17, sendo que quase metade dos bairros nao apresenta área verde. Por outro lado, alguns bairros apresentam uma proporção superior a 0.5, indicando que a área com cobertura natural é maior do que a área ocupada, com destaque para o Alto da Boa Vista. Por fim, a temperatura noturna do solo média é 20.28°C, com uma baixa variabilidade entre os bairros.

O modelo ajustado avaliou o risco de arboviroses e a probabilidade de ocorrência da chikungunya em comparação a dengue e Zika. Essa escolha foi feita devido a maior ocorrência de casos de chikungunya na cidade do que as outras duas arboviroses.

A Figura 2 apresenta os riscos de arboviroses (em escala logarítmica). Note que há clara mudança do nível do risco em bairros vizinhos. O modelo ajustado identifica as possíveis fronteiras que indicam descontinuidade na superfície de risco sinalizadas em linhas pretas. A maioria dos limites estimados corresponde às mudanças na superfície de risco, sugerindo que as covariáveis parecem ser boas métricas de dissimilaridade para detectar descontinuidades. O mesmo pode ser visto na Figura 3, que apresenta as probabilidades de ocorrência de chikungunya. Note que Manguinhos, bairro formado exclusivamente por áreas favelizadas de baixa renda e com alta probabilidade de ocorrência de chikungunya em relação as outras arboviroses analisadas, possui quebras de contiguidade com quase todos os seus vizinhos, exceto o bairro do Jacarezinho, que possui características

territoriais similares. O Complexo do Alemão também apresentou descontinuidade com Olaria, Ramos e Higienópolis. Outro bairro que chama atenção é o Alto da Boa Vista, que possui alto risco para arboviroses. Este bairro apresenta descontinuidade com todos os seus vizinhos limítrofes. Observe também que Jacarepaguá possui quebras com bairros vizinhos que possuem características socioeconômicas diferentes (Anil, Curicica, Cidade de Deus, Taquara, Realengo, Gardênia Azul, Água Santa, Engenho de Dentro e Lins de Vasconcelos). Além disso, vale ressaltar que o risco estimado para arboviroses desses vizinhos é mais elevado.

	α_1	α_2	α_3	α_4	α_5	τ	τ_{uv}	a	μ
Média	0.278	0.058	0.143	0.066	0.063	2.266	1.615	0.018	-0.179
DP	0.312	0.047	0.084	0.051	0.049	0.938	0.150	0.017	0.151
2.5%	0.005	0.005	0.018	0.005	0.003	1.059	1.340	0.000	-0.464
50%	0.155	0.046	0.131	0.054	0.053	2.064	1.608	0.013	-0.182
97.5%	1.060	0.181	0.358	0.179	0.190	4.735	1.919	0.057	0.097
α^{min}	0.088	0.141	0.114	0.142	0.133	-	-	-	-

Tabela 2: Média, mediana, desvio padrão e os quantis de 2.5% e 97.5% a *posteriori* de alguns parâmetros do modelo.

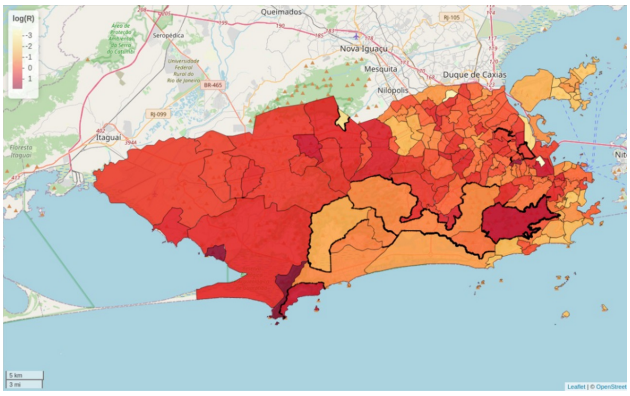


Figura 2: Mapa com o log dos riscos estimados e os limites de descontinuidades detectados. Linha preta: indica descontinuidade.

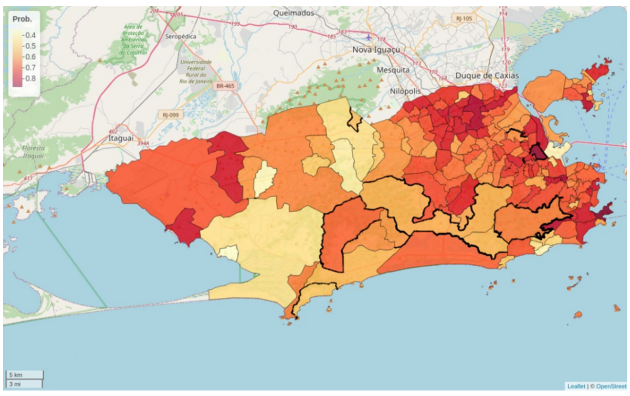


Figura 3: Mapa com as probabilidades de ocorrência da chikungunya estimadas e os limites de descontinuidades detectados. Linha preta: indica descontinuidade.

Espera-se que os resultados apresentados neste trabalho contribuam para a discussão da dinâmica da dengue, Zika e chikungunya na cidade do Rio de Janeiro, buscando identificar áreas de maior risco de transmissão para que haja ações focalizadas dos agentes públicos.

Referências

[1] Adin, A., Lee, D., Goicoa, T. e Ugarte, M. A two-stage approach to estimate spatial and spatio-temporal disease risks in the presence of local discontinuities and clusters. *Stat Methods Med Res 28* (2019), 2595–2613.

[2] Almeida, A. S., Medronho, R. A. e Valencia, L. I. O. Spatial analysis of dengue and the socioeconomic context of the city of rio de janeiro (southeastern brasil).

[3] Anderson, C., Lee, D. e Dean, N. Identifying clusters in bayesian disease mapping. *Biostatistics 28* (2014), 1–13.

[4] Besag, J., York, J. e Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics 43*, 1 (1991), 1–20.

[5] Desjardins, M. R., Eastin, M. D., Paul, R., Casas, I. e Delmelle, E. M. Space-time conditional autoregressive modeling to estimate neighborhood-level risks for dengue fever in cali, colombia. *Am J Trop Med Hyg 103(5)* (2020), 2040–2053.

[6] dos Santos, J. P. C., Honório, N. A. e Nobre, A. A. Definition of persistent areas with increased dengue risk by detecting clusters in populations with differing mobility and immunity in rio de janeiro, brazil.

[7] Freitas, L. P., Schmidt, A. M., Cossich, W., Cruz, O. G. e Carvalho, M. S. Spatio-temporal modelling of the first chikungunya epidemic in an intra-urban setting: The role of socioeconomic status, environment and temperature. *PLoS Negl Trop Dis 15(6): e0009537* (2021).

[8] Honório, N. A., Câmara, D. C. P., Calvet, G. A. e Brasil, P. Chikungunya: an arbovírus infection in the process of establishment and expansion in brazil. *Cad Saúde Pública 31 (5)* (2015), 906–8.

[9] Illian, J. B., Martino, S., Sørbye, S. H., Gallego-Fernández, J. B., Zunzunegui, M., Esquivias, M. P. e Travis, J. M. Fitting complex ecological point process models with integrated nested laplace approximation. *Methods in Ecology and Evolution 4*, 4 (2013), 305–315.

[10] Lee, D. e Mitchell, R. Boundary detection in disease mapping studies. *Biostatistics 13*, 3 (2012), 415–426.

[11] Leroux, B. G., Lei, X. e Breslow, N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 179–191.

[12] Lima-Camara, T. N. Arboviroses emergentes e novos desafios para a saúde pública no brasil. *Revista de Saúde Pública 50:36* (2016).

[13] Reisen, W. Landscape epidemiology of vector-borne diseases. *Annu Rev Entomol* (2010), 461–483.

[14] Santos, J. P. C., Honório, N. A., Barcellos, C. e Nobre, A. A. A perspective on inhabited urban space: Land use and occupation, heat islands, and precarious urbanization as determinants of territorial receptivity to dengue in the city of rio de janeiro.

[15] Teixeira, T. R. A. e Cruz, O. G. Spatial modeling of dengue and socioenvironmental indicators in the city of rio de janeiro, brazil. *Cad. Saúde Pública 27(3)* (2011), 591–602.

Avaliação da pobreza no estado do Rio de Janeiro: o impacto da formalidade

Marcson Araújo (UFF)
Rafael Erbisti (UFF)
Carolina Botelho (Mackenzie)

Email de contato: azevedomarcson@id.uff.br, rerbisti@id.uff.br, carolinabotelhomch@gmail.com.

Resumo

Este trabalho busca observar fatores de indivíduos e de seus domicílios associados à condição de pobreza monetária na Região Metropolitana do Rio de Janeiro no último trimestre de 2020. A partir dos dados da PNAD Contínua, verifica-se a condição socioeconômica da população durante o recente período atípico, causado pela pandemia da COVID-19, e avaliar os efeitos de fatores relacionados à pobreza monetária. Para estimar a influência destes fatores, foi utilizado o modelo logístico para dados binários sob a ótica bayesiana. Nos resultados, observou-se que a condição de ocupação formal indica aumento nas chances de estar fora da linha de pobreza.

Palavras-chave: Pobreza. Formalidade. Modelo Linear Generalizado. Modelo Logit. Inferência Bayesiana.

Introdução

Próximo a virada do século nos anos 1900 foram divulgadas as primeiras publicações sobre pobreza, realizadas na Inglaterra, onde foi identificada a importância da regionalização, encontrando diferentes valores para ser considerado pobre entre bairros londrinos [1]. Uma abordagem que mudou o curso da classificação de pobreza no mundo, foi vista pela primeira vez no estudo da vida na cidade em 1901 [8]. Neste estudo, foi considerado pobre quem não gerava renda semanal igual ou acima do valor para adquirir uma cesta de alimentos (medida em valor calórico) para que um adulto mantenha o peso.

No Brasil, a partir de 1990, a produção acadêmica sobre pobreza ganhou relevância, principalmente após a implementação do Plano Real, que abriu portas para a questão ser amplamente discutida no país. Nessa época, era cada vez mais perceptível que muitos brasileiros não tinham condições de ter acesso a itens básicos para consumo e sobrevivência e, então, a redistribuição de riqueza começou a ser pensada como um contorno que ampliaria os direitos sociais e permitiria uma melhora direta na renda familiar. Nesse sentido, a expectativa dos especialistas era que houvesse geração de capital humano e, em algum momento, não fosse mais preciso a família participar de programas de redistribuição de renda [7]. Houve, portanto, uma atualização de programas de transferência de renda que eram pouco conhecidos e para poucas pessoas (idosos, portadores de deficiência pobres e focados em famílias pobres com crianças). Em 2003, houve um processo de consolidação de diversos programas de transferência de renda, gerando um único programa de redistribuição de renda, conhecido como Programa Bolsa Família (PBF). [7]

A renda está associada diretamente ao trabalho exercido pelo indivíduo e, com isso, um mercado de trabalho volátil, com altos percentuais de contratações e de demissões, indica alta variação na renda dos mais pobres no decorrer dos meses. Segundo o CAGED, no Rio de Janeiro, em 2020[1], o saldo de empregos formais foi negativo, de -133.754 vagas, pior Unidade da Federação nos dados apresentados pelo Ministério da Economia [2]. A variação percentual do saldo em relação às admissões foi de -16%. Neste contexto, deseja-se medir o impacto da formalidade na condição de pobreza de indivíduos que residem em domicílios com pelo menos um dos moradores ocupados no mercado de trabalho formal e analisar as diferenças nas características socioeconômicas dos residentes pobres na Região Metropolitana e Capital do estado do Rio de Janeiro.

[1] Janeiro a Novembro de 2020.

Material e métodos

A fonte de informações para a realização do trabalho é uma pesquisa domiciliar com grande capilaridade no Brasil, a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), que consulta todo o país anualmente, exceto em anos de Censo. A pesquisa é implementada em visitas, cinco no total, apuradas a cada trimestre. As perguntas compreendem os temas sobre trabalho, educação e assistência, além de informações básicas do domicílio e de seus moradores[2]. As principais informações utilizadas neste trabalho, contemplam renda domiciliar *per capita*, condição de ocupação no domicílio, escolaridade do responsável do domicílio, localização do domicílio, sexo e cor. Os dados avaliados pertencem à base do quarto trimestre de 2020.

Entre o conjunto de variáveis derivadas da pesquisa, a renda domiciliar *per capita* permite classificar os indivíduos pobres através do ponto de corte da linha de pobreza de elegibilidade do PBF, que no período da pesquisa era de R$178,00.

De acordo com os valores possíveis para a variável de interesse composta pelos eventos de, o indivíduo ser pobre ou não, existe um modelo estatístico que é capaz de estimar valores a este tipo de experimento, que pode ser dito um ensaio de Bernoulli. É atribuído então a variável de interesse, Y_i, $i = 1, \ldots, n$, a distribuição de Bernoulli que pertence à família exponencial de distribuições e permite o uso dos Modelos Lineares Generalizados (MLG) [3]. Dado que o valor esperado da distribuição de Y_i é a probabilidade π_i, o MLG proposto tem a capacidade de estimar a probabilidade de um indivíduo estar abaixo da linha de pobreza a partir de características individuais e domiciliares. Dessa forma, temos que $\pi_i = P(Y_i = 1)$, a probabilidade de sucesso para a condição de pobreza para o indivíduo i, $i = 1, \ldots, n$. O modelo linear generalizado geral pode ser descrito da seguinte forma

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \quad i = 1, \ldots, n \\ ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \mathbf{X}_i'\boldsymbol{\beta} \end{aligned} \tag{1}$$

onde $\mathbf{X}_i'$ é a matriz de covariáveis que medem características individuais e domiciliares utilizadas para explicar a classificação do indivíduo como pobre ou não pobre, de dimensões $n \times p$; $\boldsymbol{\beta}$ é o vetor de parâmetros que medem os efeitos das características dos indivíduos na probabilidade de estar abaixo da linha de pobreza, de dimensão $p \times 1$.

As variáveis explicativas usadas na equação (2) são as informações socieconômicas do indivíduo e do domicílio e a nomenclatura utilizada está representada na Tabela 1. Portanto, considerando os fatores de interesse listados na Tabela 1, o modelo da equação (1) pode ser reescrito como

$$\begin{aligned} Y_i &\sim \text{Bernoulli}\,(\pi_i) \quad i = 1, \ldots, n \\ ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1 \times area_i + \beta_2 \times sexo_i + \beta_3 \times cor_i + \\ &\quad \beta_4 \times eresp2_i + \beta_5 \times eresp3_i + \beta_6 \times eresp4_i + \beta_7 \times eresp5_i + \\ &\quad \beta_8 \times ocupacaoSF_i + \beta_9 \times ocupacaoFI_i \end{aligned} \tag{2}$$

Sob a abordagem bayesiana, para especificação completa do modelo, é preciso definir as distribuições a *priori* dos parâmetros desconhecidos. Neste caso, é preciso especificar as *prioris* para o vetor $\boldsymbol{\beta}$. Assim, assume-se que $\boldsymbol{\beta} \sim \text{NMV}\,(\mathbf{0}, V_\beta)$, com $V_\beta = 1000 I_p$, I_p sendo a matriz identidade de dimensão $p \times p$.

Para o processo de estimação dos parâmetros desconhecidos, foram realizadas simulações da distribuição a *posteriori* através de métodos iterativos de Monte Carlo via Cadeias de Markov (MCMC). A implementação foi feita no *software* `R` [6] a partir do pacote `rjags` [5]. A análise de convergência foi realizada a partir dos métodos disponíveis no pacote *coda* [4].

Resultados e discussão

Com os indicadores observados, explora-se o conjunto de dados para identificar características dos indivíduos. A avaliação inicial mostra um percentual de pobres de 3,8% na Capital e de 5,1% na Região Metropolitana. Destacam-se desequilíbrios na população alvo como, para a dimensão de

[2] Página oficial PNAD Contínua.

Variável	Identificação	Categorias	Referência
area	Tipo de área	Região Metropolitana	*
		Capital	
sexo	Sexo	Mulher	*
		Homem	
cor	Cor ou Raça	Pretos ou Pardos	*
		Brancos	
eresp	Escolaridade do Responsável do Domicílio	1 Sem Instrução	*
		2 Fundamental Incompleto	
		3 Fundamental Completo	
		4 Médio Completo	
		5 Superior Completo	
ocupacao	Condição de Ocupação no Domicílio	SI Somente Informal	*
		SF Somente Formal	
		FI Formal e Informal	

Tabela 1: Exemplo do resultado da estimação do modelo logístico

cor ou raça, na RM, 61,5% são pretos ou pardos e na Capital este percentual é de 44,0%. Para a distribuição da escolaridade do responsável, na Capital, 78,7% dos responsáveis dos domicílios completaram o ensino médio (categorias 4 e 5). Para a RM, esse percentual é de 61,4%. O percentual de ocupados entre toda a população alvo é de 51,1%. Ao realizar cruzamentos observa-se que dentre os indivíduos classificados abaixo da linha de pobreza, o percentual de ocupados é de 32,9%. No grupo de mulheres, menos da metade, 44,3%, está ocupada.

De posse das informações, para a variável de condição de ocupação no domicílio na Figura 1, identifica-se a presença da informalidade entre os indivíduos abaixo e acima da linha de pobreza. Observe que, de forma geral, tanto na Capital quanto na RM, a maioria dos indivíduos que está acima da linha de pobreza ou vive em domicílio em que todos que trabalham são formais ou tem, pelo menos, um trabalhador formal. Por outro lado, dentre os indivíduos pobres, mais de 72% vivem em domicílios na Capital em que todos os trabalhadores são informais. Na Região Metropolitana, esse número é de 84%.

Após o ajuste do modelo e aplicação dos testes de convergência das cadeias simuladas, pode-se avaliar os resultados encontrados e apresentados na Tabela 2.

Coeficientes	Média a *posteriori*	Desvio Padrão a *posteriori*	Razão de chance	2.5%	97.5%	Probabilidade de significância
Intercepto	-0.90	0.36	-	-	-	1.00
area	0.09	0.10	1.094	0.896	1.323	0.81
sexo	-0.05	0.10	0.951	0.787	1.150	0.70
cor	-0.07	0.11	0.932	0.756	1.150	0.75
eresp2	-0.33	0.31	0.719	0.399	1.363	0.86
eresp3	-1.06	0.33	0.346	0.186	0.670	1.00
eresp4	-0.92	0.31	0.399	0.219	0.748	1.00
eresp5	-1.44	0.34	0.237	0.125	0.472	1.00
ocupacaoSF	-2.48	0.13	0.084	0.063	0.108	1.00
ocupacaoFI	-3.28	0.30	0.038	0.020	0.066	1.00

Tabela 2: Estimativas a *posteriori* dos parâmetros do modelo.

Note que apesar dos intervalos de credibilidade das razões de chance das variáveis *area*, *sexo*,

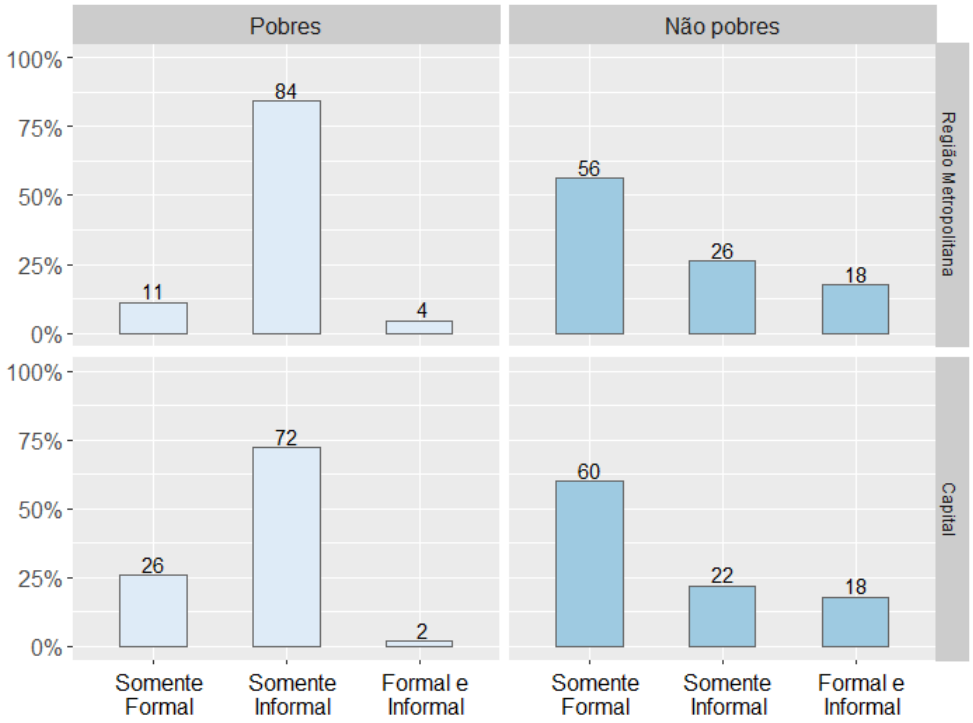


Figura 1: Gráfico da distribuição de pessoas dado a condição de ocupação no domicílio segundo o ano e tipo de área

cor e *eresp2* conterem o valor unitário, há alta probabilidade de significância. Isso significa que a massa de probabilidade da distribuição a *posteriori* indica efeito significativo dessas variáveis. Por exemplo, podemos interpretar a variável *area* da seguinte forma: residir na capital aumenta a chance de um indivíduos estar abaixo da linha de pobreza em 9,4% em comparação a um individuo que reside na RM, com probabilidade de 0,81 desse efeito ser, de fato, positivo.

Os valores da razão de chances para os efeitos de cada fator é dita pela relação entre as chances de ser pobre dado uma condição observada e não observada. Assim, temos que, comparada a indivíduos em domicílios com o responsável sem instrução (*eresp1*), um indivíduo que reside em um domicílio cuja a escolaridade do responsável é ensino fundamental completo (*eresp3*) tem chances 65,4% menor de ser pobre. Seguindo esta forma de análise, quando a escolaridade do responsável do domicílio é ensino superior completo (*eresp5*), a chance do indivíduo ser pobre é 76,5% menor do que um indivíduo que reside num domicílio cujo responsável é analfabeto.

Nas categorias de ocupação no domicílio, a razão de chances é comparada ao domicílio com somente ocupados informais (*ocupacaoSI*). Logo indivíduos que residem em domicílios com somente ocupados formais (*ocupacaoSF*) têm chance 91,6% menor de estarem abaixo da linha de pobreza. Quando o indivíduo reside em domicílio em que há pessoas ocupadas formais e informais (*ocupacaoFI*) a chance de ser pobre é 96,2% menor do que indivíduos que vivem em domicílios em que todos os ocupados são informais. Portando, a condição de ocupação de ao menos um ocupado no mercado de trabalho formal dentro do domicílio aumenta substancialmente as chances de um indivíduos não estar abaixo da linha de pobreza.

Espera-se que os resultados obtidos neste trabalho possam subsidiar a criação de políticas públicas para aumentar a formalidade dos trabalhadores no estado do Rio de Janeiro, auxiliando a redução de pobreza.

Referências

[1] Booth, C. Life and labour of people in london.

[2] de Previdência e Trabalho. Ministério da Economia, S. E. Apresentação de estatísticas mensais do emprego formal - novo caged.

[3] Nelder, J. A. e Wedderburn, R. W. M. Generalized linear models.

[4] Plummer, M., Best, N., Cowles, K. e Vines, K. Coda: Convergence diagnosis and output analysis for mcmc. *R News 6*, 1 (2006), 7–11.

[5] Plummer, M., Stukalov, A. e Denwood, M. *Pacote rjags, Bayesian Graphical Models using MCMC*. 2019.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[7] Rocha, S. *Tranferências de renda no Brasil. O fim da Pobreza?* Alta Books, 2019.

[8] Rowntree, S. Poverty: A study of a town life.

Modelos espaço-temporais para dados de contagem

Matheus Alves Pereira dos Santos (UFF)
Jony Arrais Pinto Junior (UFF)

Email de contato: matheusapds@id.uff.br.

Resumo

Este trabalho discute o estudo de fenômenos com dependência espaço-temporal que podem ser descritos por meio de dados de contagem. O modelo analisado baseia-se no modelo espacial proposto por Leroux et al. (2000) [6], enquanto a dependência temporal é considerada por meio de um caso particular de modelos lineares dinâmicos generalizados, o polinomial de primeira ordem. Toda a inferência utilizada neste trabalho baseia-se na teoria de inferência Bayesiana e na utilização do método de amostragem de Monte Carlo Hamiltoniano, por meio do software Stan. A avaliação do modelo proposto foi realizada por meio de um estudo de simulação, em que se considerou cinco diferentes cenários variando-se as condições de dependência espacial e temporal, assim como a variabilidade dos dados. Os resultados deste processo simulado revelaram um desempenho satisfatório do modelo no que diz respeito à estimação dos parâmetros de interesse em todos os cenários contemplados.

Palavras-chave: Modelo espaço-temporal. Dados de contagem. CAR. Inferência Bayesiana.

Introdução

Dados agregados por regiões no espaço são comumente o objeto de interesse de pesquisadores de diversas áreas. Pode-se citar como exemplo o número de casos de pessoas infectadas com COVID-19 nos bairros de Niterói ou ainda o número de homicídios nos municípios do estado do Rio de Janeiro. Fenômenos deste tipo, usualmente, apresentam uma estrutura de dependência no espaço.

O desafio de analisar dados com essas características, levando-se em consideração as particularidades inerentes à cada área de estudo, fez com que diversas metodologias fossem desenvolvidas ao longo do tempo. Dentre essas diferentes abordagens, para trabalhar com dados de contagens agregados, tem-se as propostas de Besag et al. (1991) [3], Besag e Kooperberg (1995) [2] e Leroux et al. (2000) [6].

Aplicações envolvendo esse tipo de dados, muitas das vezes, podem facilmente ser expandidas para problemas em que existe algum tipo de dependência temporal. Para isso, basta que tenha-se interesse, por exemplo, em modelar o número de homicídios semanais nos municípios do Rio de Janeiro no ano de 2020. Nesses casos, diz-se que os dados apresentam uma estrutura espaço-temporal. Devido a sua grande aplicabilidade em diferentes áreas do conhecimento, das quais destacam-se a epidemiologia, segurança pública e medicina, metodologias apropriadas para lidar com diferentes formas de estruturação espaço-temporal mostram-se de suma importância.

Dentro da literatura existem diferentes abordagens que têm por objetivo modelar dados com esse tipo de dependência, das quais pode-se citar os trabalhos de Bernardinelli et al. (1995) [1] e Knorr-Held e Besag (1998) [5]. Embora sejam abordagens já consolidadas, elas apresentam características que acabam por limitar a sua aplicabilidade. O modelo proposto por Bernardinelli et al. (1995) [1], por exemplo, faz a suposição de normalidade dos dados além da linearidade dos efeitos temporais.

Diante do exposto, é inegável a importância da compreensão das dependências espaciais e/ou temporais para inúmeras áreas, principalmente, quando se trabalha com dados agregados de contagens. Assim sendo, este trabalho, que será realizado à luz da teoria Bayesiana, tem como objetivo principal a proposição e o estudo de um modelo espaço-temporal de dados agregados de contagem baseados na distribuição Poisson.

A definição deste modelo ocorrerá de forma semelhante à realizada por Vivar (2007) [9], embora aqui, a dependência espacial será descrita por meio da distribuição normal proposta por Leroux et al. [6], devido à sua maior flexibilidade. Já o efeito temporal será estruturado como um caso

particular de um modelo dinâmico, o polinomial de primeira ordem. Esta escolha repousa no resultado mostrado por Vivar (2007) [9], em que, esta abordagem, embora fosse a mais simples, obteve um dos melhores desempenhos dentre as formas de estruturação temporal testadas. Além disso, a forma como se definirá os efeitos espaciais e temporais no modelo proposto baseia-se no argumento de Waller et al. [10], que afirma que a escolha da utilização de uma interação entre esses efeitos, embora tenha potencial para melhorar a qualidade dos ajustes, deve ser feita com cuidado. Essa necessidade de uma maior atenção à esses efeitos espaço-temporais reside na complexidade de se conciliar a escala espacial com a temporal. Além disso, à utilização dessas interações, normalmente, aumenta a complexidade do modelo, uma vez que, o número de parâmetros a serem estimados cresce. Visando evitar essas possíveis problemáticas, optou-se por considerar, no modelo proposto, os efeitos espaciais e temporais de forma aditiva e independente.

Por fim, a avaliação do desempenho do modelo na estimação dos parâmetros, será realizada por meio de um estudo de simulação. Visando tornar mais amplo o entendimento da capacidade dessa metodologia, serão considerados diferentes cenários de dependência espacial, de memória dos efeitos temporais, assim como de variabilidade dos dados.

Material e métodos

Antes de iniciar, de fato, a análise espaço-temporal de dados agregados de contagem, fez-se necessário definir o contexto ao qual o modelo será descrito. Para isso, considere uma região de interesse A particionada em S sub-regiões para as quais foram observadas realizações de uma variável de interesse Y, que representa contagens da ocorrência de um determinado evento, para T intervalos de tempo. Sejam Y_{st} e e_{st}, respectivamente, a contagem observada e o número de observações esperado de uma determinada variável de interesse na sub-região s, no tempo t. Por fim, considere $\mathbf{X}$, como a matriz de covariáveis.

Diante do cenário descrito, tem-se que a definição matemática do modelo espaço-temporal proposto é dada por:

$$
\begin{aligned}
Y_{st}|\lambda_{st}e_{st} &\sim Poi(\lambda_{st}e_{st}), \quad s = 1, \cdots, S; \quad t = 1, \cdots, T, \\
log(\lambda_{st}) &= \mathbf{X}\boldsymbol{\beta} + \epsilon_s + \psi_t, \quad s = 1, \cdots, S; \quad t = 1 \cdots, T, \\
\epsilon_s|\boldsymbol{\epsilon}_{-s} &\sim N\left(0, \frac{1}{\tau_\epsilon(\rho\sum_{j=1}^{S} w_{sj} + 1 - \rho)}\right), \\
\psi_t &= \phi\psi_{(t-1)} + d_t, \quad t = 2, \cdots, T, \\
d_t &\sim N\left(0, \frac{1}{\tau_\psi}\right), \\
\boldsymbol{\beta} &\sim NM_p(\mathbf{A}_\beta, \mathbf{V}_\beta), \\
\phi &\sim N(a_\phi, v_\phi), \\
\tau_\psi &\sim Gama(a_\psi, b_\psi), \\
\tau_\epsilon &\sim Gama(a_\epsilon, b_\epsilon), \\
\rho &\sim Unif(0, 1),
\end{aligned}
\tag{1}
$$

em que, λ_{st} é o risco associado à variável de interesse na sub-região s, no período de tempo t. Já $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \cdots, \mathbf{x}_p)$ é o vetor de covariáveis que podem variar no espaço, no tempo, ou ainda em ambas as dimensões e $\boldsymbol{\beta}$ é o vetor p variado de efeitos associados a cada uma dessas variáveis explicativas ao qual é atribuído uma distribuição Normal p variada com um vetor de médias $\mathbf{A}_\beta$ e matriz de variâncias e covariâncias $\mathbf{V}_\beta$. $\boldsymbol{\epsilon}$ é o vetor de efeitos aleatórios estruturado no espaço, em que, essa dependência é modelada por meio da distribuição normal proposta por Leroux et al. [6]. A média dessa distribuição é definida como zero, seguindo a proposta de Rampaso (2014) [7]. Já a variância é inversamente proporcional ao número de vizinhos que a sub-região s possui. Esse esquema de vizinhança é definido na matriz de proximidade espacial $\mathbf{W}$ em que o elemento w_{sj} diz se as sub-regiões s e j compartilham ou não fronteiras. Por fim, tem-se que τ_ϵ é o parâmetro de precisão espacial e ρ é a autocorrelação global. Aqui, é válido destacar que a inclusão dessa autocorrelação como um hiperparâmetro tornou o modelo CAR Leroux [6] mais flexível em relação à intensidade dessa autocorrelação global. Já ao vetor de efeitos temporais $\boldsymbol{\psi}$ é atribuído o modelo dinâmico polinomial de primeira ordem que, na verdade, trata-se de um autorregressivo de ordem

1. ϕ é definido como o parâmetro de memória dos efeitos e indica a estacionariedade do modelo, caso esteja no intervalo $[-1, 1]$. Já τ_ψ é a precisão dos efeitos temporais enquanto d_t é um termo de erro aleatório normalmente distribuído.

Uma vez definido o modelo, tem-se que o vetor de parâmetros de interesse é dado por $\boldsymbol{\theta} = (\boldsymbol{\epsilon}, \rho, \tau_\epsilon, \boldsymbol{\psi}, \phi, \tau_\psi, \boldsymbol{\beta})^T$. A especificação deste vetor de parâmetros é crucial, pois é a partir da estimação de seus elementos que se dará o processo de inferência do modelo. Essa etapa inferencial pode ser feita de algumas maneiras distintas. Neste trabalho, porém, optou-se por recorrer à abordagem Bayesiana para a realização de todo processo inferencial do modelo proposto.

A inferência Bayesiana busca mensurar incertezas sobre quantidades não observadas por meio de uma combinação de toda a informação subjetiva relacionada a um problema, resumida nas distribuições a *priori*, e o conhecimento proveniente dos dados observados, por meio da função de verossimilhança. Toda essa informação é combinada na distribuição a *posteriori*, por meio do Teorema de Bayes, que utiliza a seguinte relação:

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto L(\boldsymbol{\theta}|\mathbf{Y})\pi(\boldsymbol{\theta}), \tag{2}$$

em que $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^T$ é o vetor formado por todos os parâmetros de interesse, $p(\boldsymbol{\theta}|\mathbf{Y})$ é a distribuição conjunta a *posteriori*, $L(\boldsymbol{\theta}|\mathbf{Y})$ é a função de verossimilhança e, por fim, $\pi(\boldsymbol{\theta})$ é a distribuição à priori conjunta, normalmente, obtida supondo a independência a *priori* dos parâmetros.

Embora a ideia seja realizar inferência a partir desta distribuição a *posteriori* ela, comumente, não apresenta uma forma fechada, o que impossibilita ou torna muito complicado o processo inferencial. Essa questão foi contornada com os avanços de métodos computacionais, principalmente ligados ao método de simulação de Monte Carlo via Cadeias de Markov (MCMC). A ideia por trás da utilização do método de MCMC é simular um experimento com o intuito de determinar as propriedades probabilísticas de uma população por meio de uma amostra aleatória de seus componentes.

Dois dos métodos de Monte Carlo via Cadeias de Markov mais utilizados são o Amostrador de Gibbs e o Algoritmo de Metropolis-Hastings. Informações mais detalhadas sobre ambos métodos podem ser encontradas em Gamerman e Lopes (2006) [4]. Embora essas metodologias sejam amplamente utilizadas, o ajuste de modelos mais complexos, como os modelos Hierárquicos Bayesianos, torna-se computacionalmente custoso, devido à forma que essas abordagem percorrem o espaço paramétrico de interesse.

Visando resolver essa questão, caso todos os parâmetros de interesse sejam contínuos, pode-se utilizar, como alternativa ao Amostrador de Gibbs e ao Algoritmo de Metropolis-Hasting, um método chamado de Monte Carlo Hamiltoniano (HMC). Esta metodologia utiliza equações hamiltonianas para realizar uma transformação do espaço paramétrico de interesse, possibilitando que seu algoritmo o consiga percorrer de uma forma mais eficiente. Essa modificação faz com que as cadeias geradas pelo HMC tendam a convergir mais rápido e a ser menos autocorrelacionadas. Devido à limitação de páginas deste trabalho, não será realizada uma discussão mais profunda sobre o funcionamento desse método, porém, essas informações podem ser encontradas em Stan Development Team [8].

Uma vez entendido o processo de inferência e estimação dos parâmetros utilizados na simulação, faz-se necessário avaliar o desempenho do modelo proposto no que tange a recuperação dos parâmetros. Para essa avaliação serão utilizadas métricas que buscam mensurar a eficiência das estimações pontuais e intervalares realizadas pelo modelo. Como medida para verificar o quão próxima a estimação pontual está do verdadeiro valor do parâmetro, foi adotado a raiz do erro quadrático médio. A definição matemática dessa medida é dada por $RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\theta-\hat{\theta}_i)^2}{N}}$, em que $\hat{\theta}_i$ é a estimativa pontual para o parâmetro de interesse θ e N é o número de simulações realizadas.

Já em relação às medidas utilizadas para a avaliação das estimações intervalares, computou-se, a cobertura do intervalo de credibilidade para os hiperparâmetros associados aos efeitos espaciais e temporais, assim como para os efeitos das covariáveis. É válido destacar que, entende-se como cobertura a porcentagem de vezes que o intervalo de credibilidade conteve o valor verdadeiro do parâmetro de interesse. Para complementar a avaliação sobre a qualidade das estimações intervalares, analisou-se também a amplitude desses intervalos, em que essa amplitude é definida como a diferença entre o limite superior e inferior do intervalo de credibilidade.

Resultados e discussão

O principal objetivo deste trabalho é a proposição de um modelo espaço-temporal para dados de contagens e, obviamente, a verificação da capacidade desta metodologia em realizar boas estimações. Para alcançar esse propósito, de forma empírica, recorreu-se a realização de um estudo de simulação. A ideia por trás deste processo é expor o modelo proposto, de forma repetida, a diferentes cenários que podem ser encontrados em aplicações reais. A partir de todos esses ajustes, pode-se avaliar o desempenho que o modelo tende a apresentar em cada um dos cenários considerados.

O estudo de simulação aqui realizado utilizou como plano de fundo os $S = 52$ bairros da cidade de Niterói, considerando $T = 30$ intervalos de tempo. As contagens Y_{st} foram geradas a partir do modelo proposto na Equação 1. Primeiramente, o número de observações esperados $\mathbf{e}_{st}$ foi gerado a partir de uma distribuição uniforme no intervalo $[20, 100]$. É importante destacar que fez-se a suposição que esses valores variavam apenas no espaço.

Em relação à geração da matriz $\mathbf{X}$, considerou-se uma única variável explicativa tal que $X \sim N(0, 1)$ em que, novamente, foi suposto que X variava apenas no espaço. Já ao efeito associado a X atribuiu-se o valor 0,5. Aqui, faz-se necessário dizer que optou-se por considerar o intercepto, $\beta_0 = 0$.

Na geração dos efeitos espaciais e temporais, para cada um dos seus hiperparâmetros definiu-se dois valores distintos, visando a criação de diferentes cenários. A Tabela 1 traz uma descrição dos 5 cenários simulados, assim como os valores dos hiperparâmetros utilizados. Por fim, é válido dizer que a matriz de proximidade espacial $\mathbf{W}$, foi gerada de forma binária, por meio de um método baseado em critérios de contiguidade, em que se duas sub-regiões dividem ao menos um ponto de fronteira, elas são ditas como sub-regiões vizinhas. A partir da Tabela 1 nota-se que o objetivo dessa simulação é avaliar o impacto que alterações na magnitude da autocorrelação espacial, memória e precisão dos efeitos exercem no modelo proposto. É válido destacar que, para cada cenário foram geradas $N = 100$ bases de dados distintas.

Tabela 1: Descrição dos cenários simulados

Cenário	τ_ϵ	ρ	τ_ψ	ϕ	Descrição
1	5	0,9	5	0,95	Precisões altas, autocorrelação forte e memória alta.
2	5	0,9	5	0,35	Precisões altas, autocorrelação forte e memória baixa.
3	5	0,2	5	0,95	Precisões altas, autocorrelação fraca e memória alta.
4	5	0,2	5	0,35	Precisões altas, autocorrelação fraca e memória baixa.
5	1	0,2	1	0,95	Precisões baixas, autocorrelação fraca e memória alta.

A análise dos resultados feita aqui, busca mensurar, por meio das métricas definidas anteriormente o comportamento mediano do modelo em relação às estimações pontuais e intervalares. Faz-se necessário destacar que, o estimador pontual adotado foi a mediana a *posteriori* devido a sua maior robustez em relação a valores extremos. Já para as estimações intervalares considerou-se o intervalo de credibilidade de 95%.

Tabela 2: Resultados dos hiperparâmetros dos efeitos espaciais.

Cenário	$RMSE_{\tau_\epsilon}$	$Cobertura_{\tau_\epsilon}$	$Amplitude_{\tau_\epsilon}$	$RMSE_\rho$	$Cobertura_\rho$	$Amplitude_\rho$
1	1,63	93%	4,94	0,20	95%	0,55
2	1,80	90%	4,71	0,20	91%	0,52
3	1,37	99%	5,77	0,16	98%	0,68
4	1,40	97%	5,39	0,16	96%	0,64
5	0,29	95%	1,21	0,14	96%	0,65

A análise da Tabela 2 evidencia o desempenho satisfatório do modelo em relação à estimação dos hiperparâmetros τ_ϵ e ρ em todos os cenários, apresentando, respectivamente, coberturas mínimas de 90% e 91%, ambas ocorrendo no cenário 2. Ambos hiperparâmetros apresentaram resultados semelhantes nos cenários 1 e 2 e nos cenários 3 e 4. Nos dois primeiros, τ_ϵ e ρ apresentaram valores mais altos de $RMSE$ ao se comparar com os cenários 3 e 4, além de terem mostrado uma menor cobertura. Entretanto, é válido dizer que, essa maior porcentagem de cobertura para o terceiro e quarto cenário pode ter sido influenciada pelo aumento da amplitude dos intervalos

de credibilidade que ocorreram nestes cenários. Um resultado interessante mostrado pela Tabela 2 é a redução da incerteza acerca da estimação de τ_ϵ, mostrada pela considerável diminuição do $RMSE$ e da amplitude, que ocorreu no cenário 5, justamente no qual considerou-se dados com maior variabilidade.

A Tabela 3 traz os resultados das estimações dos hiperparâmetros associados aos efeitos temporais. Avaliando esses resultados nota-se um comportamento bem semelhante ao mostrado na Tabela 2. Mais uma vez, o modelo apresentou uma menor incerteza em relação à estimação do parâmetro de precisão τ_ψ justamente no cenário em que essa precisão era menor. Já em relação ao parâmetro de memória ϕ, nota-se que o modelo apresentou estimações ligeiramente melhores para os cenários 1, 3 e 5, em que considera-se $\phi = 0,95$. Além de $RMSE$ inferior, esses cenários apresentaram coberturas próximas aos demais cenários, porém com intervalos de credibilidade com quase a metade da amplitude apresentada nos cenários 2 e 4.

Tabela 3: Resultados dos hiperparâmetros dos efeitos temporais.

Cenário	$RMSE_{\tau_\psi}$	$Cobertura_{\tau_\psi}$	$Amplitude_{\tau_\psi}$	$RMSE_\phi$	$Cobertura_\phi$	$Amplitude_\phi$
1	1,66	97%	5,22	0,15	92%	0,39
2	1,66	94%	5,12	0,17	96%	0,71
3	1,33	96%	5,25	0,13	91%	0,28
4	1,40	99%	5,65	0,18	95%	0,69
5	0,39	94%	1,09	0,14	93%	0,31

Por fim, a Tabela 4 exibe os resultados referentes à estimação do efeito da covariável, assim como o tempo computacional necessário para o ajuste do modelo. A análise destes resultados evidenciam, mais uma vez, o bom desempenho do modelo no que tange à estimação dos parâmetros e hiperparâmetros. Fazendo uma análise do efeito β associado a variável explicativa X, percebe-se que, para todos os cenários, obteve-se altas porcentagens de cobertura, sempre acima dos 90%, com intervalos de credibilidade com amplitudes relativamente pequenas. Diferentemente do observado nas Tabelas 2 e 3, é possível notar que o modelo demonstrou uma maior incerteza em relação às estimações no cenário 5, uma vez que apresentou uma menor cobertura com maiores valores de $RMSE$ e amplitude. Já em relação ao tempo computacional necessário para realizar um ajuste deste modelo, verificou-se que este não se mostrou muito custoso, levando um pouco mais de 3,5 minutos para os cenários 1, 2 e 4. Os demais cenários se destacaram com tempos mais elevados, sendo necessário quase 6 minutos para o ajuste de um modelo no último cenário.

Tabela 4: Resultados do parâmetro β e do tempo de ajuste.

Cenário	$RMSE_\beta$	$Cobertura_\beta$	$Amplitude_\beta$	$Tempo_{min}$
1	0,04	95%	0,16	3,78
2	0,04	95%	0,16	3,64
3	0,06	93%	0,21	4,69
4	0,05	96%	0,21	3,62
5	0,13	92%	0,47	5,81

O objetivo deste trabalho é propor uma nova metodologia espaço-temporal para dados agregados de contagens, que fosse, dentro das possibilidades, simples e eficiente em diferentes cenários. A importância de um método como este reside na grande aplicabilidade que dados com dependência espacial e temporal possuem. Dentre as contribuições deste trabalho para alcançar os seus objetivos, pode-se citar a estruturação dos efeitos espaciais baseando-se na proposta de Leroux et al. [6], o que garantiu maior flexibilidade ao modelo proposto. Além disso, a utilização do HMC como método de amostragem tornou esta nova metodologia mais eficiente computacionalmente.

É válido ressaltar que, devido à limitação do número de páginas deste relatório, apenas parte das análises feitas a partir do estudo de simulação foram discutidas aqui. Ainda assim, diante dos resultados mencionados aqui, pode-se concluir que o objetivo principal deste trabalho foi alcançado com sucesso uma vez que o modelo proposto mostrou estimações satisfatórias para todos os parâmetros, dentro dos 5 cenários analisados.

Referências

[1] Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. e Songini, M. Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine 14* (1995), 2433–2443.

[2] Besag, J. e Kooperberg, C. On conditional and intrinsic autoregressions. *Biometrika 84*, 4 (1995), 733–746.

[3] Besag, J., York, J. e Molllie, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43*, 1 (1991), 1–20.

[4] Gamerman, D. e Lopes, H. F. *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference.*, 2 ed. CHAPMAN and HALL, 2006.

[5] Knorr-Held, L. e Besag, J. Modelling risk from a disease in time and space. *Statistics in Medicine 17* (1998), 2045–2060.

[6] Leroux, B. G., Lei, X. e Breslow, N. *Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence.* M. Elizabeth HalloranDonald Berry, 2000, ch. 4, pp. 179–191.

[7] Rampaso, R. C. Análise bayesiana de dados espaciais explorando diferentes estruturas de variância. Mestrado tese, Universidade Estadual Paulista, 2014.

[8] Team, S. D. Stan modeling language users guide and reference manual. 2015.

[9] Vivar, J. C. *Modelos espaço-temporias para dados de área na família exponencial.* Doutorado tese, Universidade Federal do Rio de Janeiro, 2007.

[10] Waller, L. A., Carlin, B. P., Xia, H. e Gelfand, A. E. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association 92*, 438 (1997), 607–617.

Objetivos de Desenvolvimento Sustentável: É possível que o Brasil alcance as metas de saúde até 2030? Estudo regional sobre a Tuberculose

Paulo Cesar Silva Andrade dos Santos (ENCE)
Ana Carolina Soares Bertho (ENCE)
Larissa de Carvalho Alves (ENCE)

Email de contato: stylemathy@gmail.com, carolbertho@gmail.com , larissaalves.ufrj@gmail.com.

Resumo

No ano de 2015, a Cúpula das Nações Unidas estabeleceu uma agenda para o Desenvolvimento Sustentável composta por um conjunto de ações e programas. Esse pacto entre os países que compõem a Cúpula estipula o total de 17 objetivos e 169 metas que deverão ser atingidas até 2030. Nessa agenda estão previstas ações nas áreas de educação, saúde, economia, meio ambiente, igualdade de gênero, energia, saneamento básico, etc. O objetivo 3 tem como propósito "Assegurar uma vida saudável e promover o bem-estar para todos, em todas as idades". Ele contém a meta de reduzir a taxa de incidência da tuberculose por 100 mil habitantes até 2030. O objetivo do projeto é verificar, através da estimação utilizando um modelo estatístico, se, seguindo as tendências atuais, o Brasil, de forma global, ou as Grandes regiões, separadamente, irão atingir a meta que diz respeito à redução de 80% da taxa de incidência da tuberculose por 100 mil habitantes. Com tal finalidade, foi elaborada uma análise a partir de dados mensais do Sistema de Informação de Agravos de Notificação (SINAN) para o período de 2001 a 2018. A partir de um modelo Poisson dinâmico foi feita a previsão da incidência de tuberculose para o Brasil e suas regiões. Os resultados apresentados evidenciam que não há tendência de queda dos novos casos dessa doença, o que impediria o Brasil, bem como qualquer uma das Grandes regiões, de atingirem a meta definida no ODS.

Palavras-chave: ODS, tuberculose, inferência bayesiana, modelo dinâmico, modelo linear generalizado.

Introdução

Em 2015, os 193 países membros das Nações Unidas adotaram os ODS (Objetivos de Desenvolvimento Sustentável), também conhecidos como Agenda 2030. Ela é a substituta dos Objetivos de Desenvolvimento do Milênio, os quais vigoraram entre 2000 e 2015. Tal agenda é definida por um conjunto de ações, programas e diretrizes que serve como guia para o trabalho da ONU e dos países membros a fim de que se atinja o desenvolvimento sustentável.

O documento contém 17 Objetivos do Desenvolvimento Sustentável (ODS) e 169 metas, as quais deverão ser alcançadas até 2030. Os ODS englobam questões econômicas e sociais, como: redução da fome, saúde, educação, etc. .

O objetivo 3 se relaciona com a saúde e o bem estar dos cidadãos e tem como finalidade "assegurar uma vida saudável e promover o bem-estar para todas e todos, em todas as idades". O ODS 3 engloba questões ligadas a mortalidade materna; mortalidade neonatal; mortalidade por produtos químicos; controle de doenças como HIV, malária,tuberculose, hepatites virais e arboviroses e etc. .

A meta 3.3, contida no ODS 3, é *"até 2030, acabar com as epidemias de AIDS, tuberculose, malária e doenças tropicais negligenciadas, e combater a hepatite, doenças transmitidas pela água, e outras doenças transmissíveis"*. Tal meta é composta por diversos indicadores relacionados a cada uma dessas enfermidades. O indicador 3.3.8 (incidência de tuberculose por 100 mil habitantes), relacionado à tuberculose, foi o foco do presente estudo.

De acordo com a terceira edição do Caderno ODS, publicada em 2019 pelo Instituto de Pesquisa Econômica e Aplicada (IPEA), a tuberculose é uma das principais causas de morte em todo o

mundo. Segundo essa publicação, estima-se que no mundo, em 2016, tenham ocorrido 10,4 milhões de novas infecções, correspondendo a uma taxa de incidência de 140 novos casos por 100 mil habitantes. Além disso, conforme dados do Ministério da Saúde, apesar da redução de 8% no número de óbitos na última década, são notificados aproximadamente 33 casos novos por 100 mil habitantes de tuberculose por ano no país, que resultam em cerca de 4,5 mil mortes.Embora existam meios de prevenção e combate, essa doença ainda é um problema para o sistema de saúde brasileiro, pois a epidemia do HIV e a presença de bacilos resistentes tornam o cenário ainda mais complexo.

De acordo com a teoria de transição epidemiológica, todos os países do mundo estariam passando por um processo de mudança nos padrões de adoecimento e morte das populações [3]. Nesse processo, as doenças degenerativas e "produzidas pelo homem" teriam, ao longo do tempo, substituído as doenças infecciosas como principais causas de mortalidade. Com base nessa teoria, era esperado que a tuberculose já tivesse desaparecido. Porém de acordo com um estudo sobre carga de doença [6], no Brasil, a transição epidemiológica não tem ocorrido de acordo com o modelo experimentado pela maioria dos países desenvolvidos.

A tuberculose, juntamente com a malária apresentou a maior carga entre sete Doenças Tropicais Negligenciadas (DTN) estudadas [8]. Além disso, também apresentou altas taxas de mortalidade e internação quando comparada com as demais DTN e a malária. A partir da análise dos dados do Departamento de Informática do SUS e da Rede Interagencial de Informações para a Saúde, houve aumento na incidência da tuberculose na década de 1980, associada à infecção pelo HIV [1]. A realização do tratamento completo é essencial para o controle da doença e é monitorado e registrado nas bases de dados do Sistema de Informação de Agravos de Notificação (SINAN).

Sendo assim, o objetivo deste trabalho é, através da análise da taxa de incidência de tuberculose por 100 mil habitantes, verificar se o Brasil, de forma global, ou as Grandes regiões, irá atingir ou não a meta de reduzir a taxa de incidência de tuberculose em 80%, até 2030.

Metodologia

Fonte de dados

A fim de fazer uma análise da incidência de tuberculose no Brasil e Grandes regiões foram utilizadas duas fontes de dados (ambas obtidas através do DATASUS):

- Sistema de Informação de Agravos de Notificação (SINAN), onde se obteve o número de casos mensais de tuberculose para o Brasil e suas regiões.
- Projeções de população produzidas pelo Instituto Brasileiro de Geografia e Estatística (Revisão 2018) para os anos de 2001 a 2018,onde se obteve o total da população para o Brasil e Grandes regiões.

No site do DATASUS foram coletadas informações do SINAN referentes ao número de casos mensais de tuberculose no Brasil e em suas regiões, no período de 2001 a 2019 (o ano de 2019 foi separado para comparar com a previsão do modelo). Também no site do DATASUS, foram coletadas informações relacionadas à estimativa/projeção da população residente no Brasil e em suas regiões, produzida pelo IBGE nesse período. Através desses dados foi possível calcular a taxa de incidência de tuberculose por 100 mil habitantes para o Brasil e suas regiões. Foi considerada a mesma população para todos os meses de cada ano.

Modelo estatístico

Os modelos dinâmicos Bayesianos generalizados permitem a análise de dados discretos, acomodando observações na família exponencial, tais como binomial e Poisson [2]. A utilização de modelos dinâmicos Bayesianos generalizados permite uma análise das componentes e características de variabilidade da série epidemiológica. Nesta pesquisa foi usado o modelo Poisson dinâmico. Esse modelo foi analisado e comparado a outros com o objetivo de verificar as vantagens e desvantagens entre diferentes modelagens para séries temporais de contagem [4].

Tal modelo faz parte da classe dos modelos dinâmicos lineares generalizados no qual a variável resposta segue uma distribuição Poisson com parâmetro λ_t. Neste caso, para certo instante t, a

incerteza com o passar do tempo será incorporada no modelo permitindo que λ_t evolua de forma suave na estrutura temporal.

Desse modo, considere $Y_1, Y_2, \cdots, Y_T$ o número de casos da doença ao longo do tempo e p a periodicidade dos dados. Por exemplo, se $p = 12$, os dados seriam mensais, se $p = 3$, os dados seriam quadrimestrais. Nesse estudo temos $T = 216$ e $p = 12$. Supomos que as variáveis aleatórias Y_t, para $t = 1, 2, \cdots, T$, são condicionalmente independentes e seguem uma distribuição de Poisson com média λ_t, ou seja o logaritmo da média evolui no tempo e pode ser decomposto em : nível (μ_t) e sazonalidade (S_t). Além disso pode-se adicionar uma componente de tendência (β_t) na equação do nível. Assim, o modelo considerado é o seguinte :

$$Y_t|\lambda_t \sim Poisson(\lambda_t)$$

$$ln(\lambda_t) = \mu_t + S_t$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + w_{1t}, \;\; w_{1t} \sim N(0, W_1)$$

$$\beta_t = \beta_{t-1} + w_{2t}, \;\; w_{2t} \sim N(0, W_2)$$

$$S_t = -(S_{t-1} + S_{t-2} + \cdots + S_{t-p+1}) + w_{3t} \;\; , \; t > 11 \quad e \quad w_{3t} \sim N(0, W_3)$$

Assumiu-se independência a priori entre os parâmetros. A distribuição a posteriori possui uma forma desconhecida. Diante disso, para que fosse possível obter amostras da distribuição a posteriori, foi utilizado o método de Monte Carlo Via Cadeia de Markov (MCMC). A fim de garantir a convergência das cadeias rodamos 1 milhão de iterações do método MCMC, descartando as primeiras 200 mil iterações como aquecimento da cadeia, tomando de 800 em 800 para diminuir a correlação entre os valores e finalizando com uma amostra de tamanho mil.Para a execução do MCMC foi utilizado o software estatístico R [5], versão 4.1.0, e o pacote bsts [7].

Resultados e discussão

Com a finalidade de compreender o comportamento da incidência da tuberculose no Brasil e Grandes regiões entre 2001 e 2018, foi realizada uma análise descritiva. Através do gráfico das séries temporais foi possível perceber um padrão na incidência da tuberculose em determinados meses do ano e a partir disso avaliar a possibilidade de acrescentar sazonalidade ao modelo a fim de captar esse efeito. Também foi possível fazer uma comparação entre a evolução da taxa de incidência mensal por 100 mil habitantes da tuberculose apresentada pelo Brasil e grandes regiões com a meta estabelecida pelo ODS 3 para o ano de 2030.

A Figura 1 apresenta a série temporal da incidência mensal de tuberculose, no Brasil e Grandes regiões, entre 2001 e 2018. Observando-se a Figura 1 é possível perceber que o número de casos de tuberculose oscilou no período estudado. O número de casos de tuberculose, tanto no Brasil, globalmente, quanto nas Grandes regiões, individualmente, atingiu seus valores mais baixos nos anos de 2006, 2007, 2013 e 2015 e seus valores mais altos em 2002, 2017 e mais recentemente, no final de 2018. É interessante ressaltar que as Grandes regiões possuem comportamento semelhante ao do Brasil, mudando apenas o intervalo de variação do número de casos de tuberculose.

A Figura 2 apresenta a série temporal da taxa de incidência anual de tuberculose por 100 mil habitantes no Brasil e Grandes regiões, entre 2001 e 2018. Analisando a Figura 2 é possível observar que o ordenamento, visto na Figura 1, entre o Brasil e Grandes regiões sofreu alteração. A região Norte possui as maiores taxas de incidência (mesmo tendo baixo número de casos absolutos quando comparada às outras regiões), seguida pelas regiões: Sudeste,Nordeste,Sul e Centro Oeste. A posição da região Centro Oeste não se alterou, pois possui poucos casos e apresenta baixa taxa de incidência. O Brasil e a região Nordeste possuem taxas de incidência semelhantes em todo o período analisado. Tanto o Brasil quanto as Grandes regiões estão distantes da meta (na Figura, aparece tracejada antes de 2015 já que não existia) estabelecida pela Agenda 2030.

Figura 1: Incidência mensal de tuberculose - Brasil e Grandes regiões, 2001 a 2018

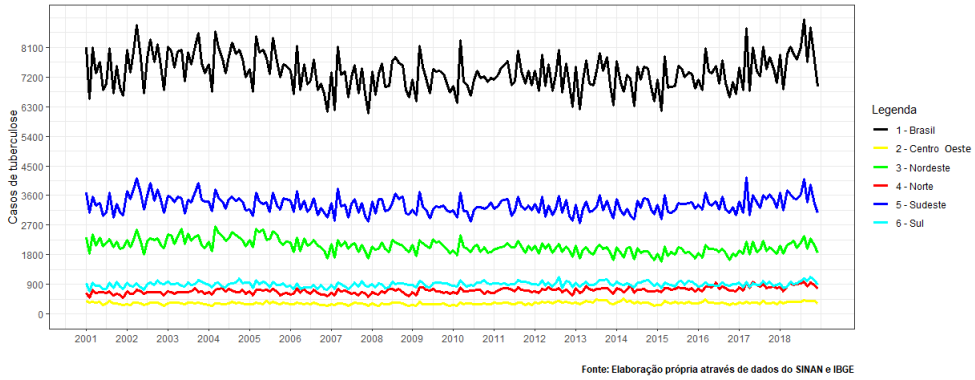
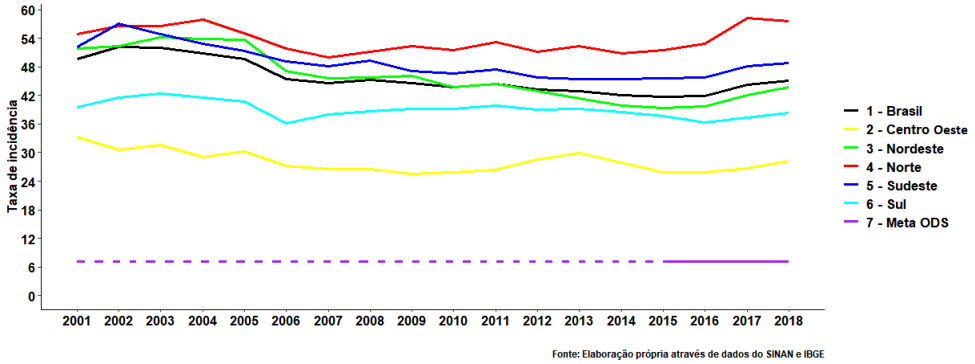
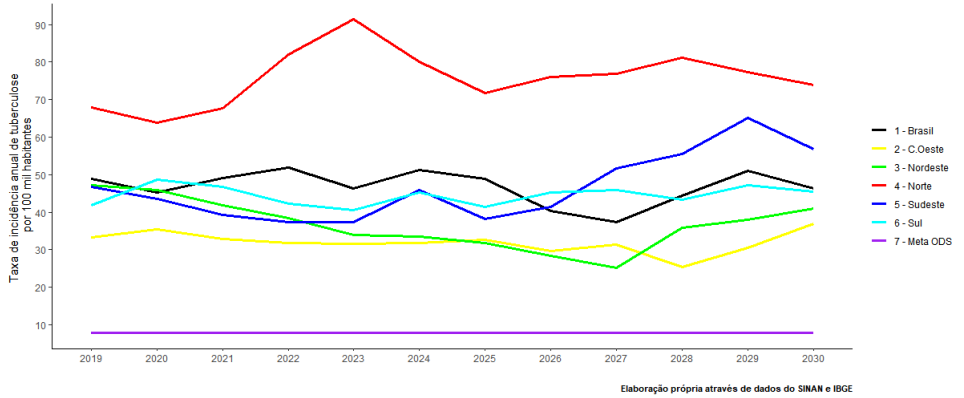


Figura 2: Taxa de incidência anual de tuberculose - Brasil e Grandes regiões, 2001 a 2018



Na Figura 3 é apresentada a previsão da taxa de incidência anual da tuberculose, por 100 mil habitantes, juntamente com a meta estabelecida pela Agenda 2030. É possível ver que a região Norte se destaca, pois têm as maiores taxas de incidência anual no período entre 2019 e 2030.Além disso percebe-se também que a região Centro Oeste, de forma geral, apresenta as menores taxas de incidência anual durante o período que realizaram-se as previsões.

Figura 3: Gráfico de previsão da taxa de incidência anual da tuberculose, por 100 mil habitantes, de 2019 até 2030 - Brasil e Grandes regiões.



Esse estudo é um desdobramento de um estudo anterior no qual foi verificado se o Brasil, de forma global, atingiria tal meta até 2030. O objetivo do estudo atual foi verificar se, seguindo o comportamento atual, o Brasil,de forma global, ou alguma das grandes regiões, de forma individual, atingirá a meta de redução da taxa incidência da tuberculose por 100 mil habitantes até 2030. Não foi identificada tendência de queda nos dados analisados (entre 2001 e 2018), pelo contrário, observou-se uma pequena tendência de crescimento, a qual é mais perceptível na previsão

a longo prazo da região Norte (região que possui, atualmente, as maiores taxas de incidência de tuberculose). Esse comportamento indica que tanto o Brasil, de forma global, quanto as Grandes regiões, de forma individual (caso a doença siga as tendências atuais) não irão atingir a meta de redução até 2030. Contudo, os intervalos de credibilidade das estimativas foram grandes, o que gera incerteza sobre as mesmas.

É importante ressaltar que o modelo estatístico utilizado (modelo dinâmico) se mostrou eficiente no contexto epidemiológico, podendo ser explorado em outros estudos na área. Esse estudo mostrou que as grandes regiões brasileiras se comportam de forma heterogênea e, além disso, estão em diferentes patamares com relação ao controle da doença no país, mostrando que algumas regiões precisam de ações públicas a fim de diminuir a taxa de incidência de tuberculose.

Mesmo que a cobertura universal e o acesso gratuito ao tratamento sejam pontos positivos no tratamento contra a tuberculose, segundo os dados disponíveis no site DATASUS, uma considerável parte da população que inicia o tratamento, cerca de 10%, acaba abandonando o tratamento, o que traz risco a si mesmo e as pessoas ao redor. Em estudos futuros poderão ser investigados os fatores associados à continuidade ou interrupção do tratamento para tuberculose, de forma a contribuir para a elaboração de estratégias de redução da incidência e da mortalidade por essa doença.

Referências

[1] Barreto, M., Teixeira, M., Bastos, F., Ximenes, R., Barata, R., et al. Sucessos e fracassos no controle de doenças infecciosas no brasil: O contexto social e ambiental, politicas, intervençoes e necessidades de pesquisa. lancet. saúde no brasil 3, 47–60. *download. thelancet. com/flatcontentassets/pdfs/brazil/brazilpor3. p df. Accessed 20* (2014).

[2] Ferreira, M. e Gamerman, D. Análise bayesiana de séries epidemiológicas de contagem via modelos dinâmicos bayesianos generalizados. *Cad Saúde Coletiva 6* (1998), 145–55.

[3] Luna, E. J. A emergência das doenças emergentes e as doenças infecciosas emergentes e reemergentes no brasil. *Revista Brasileira de Epidemiologia 5* (2002), 229–243.

[4] Pereira, J. B. Modelos para dados de contagem com estrutura temporal.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[6] Schramm, J. M. d. A., Oliveira, A. F. d., Leite, I. d. C., Valente, J. G., Gadelha, Â. M. J., Portela, M. C. e Campos, M. R. Transição epidemiológica e o estudo de carga de doença no brasil. *Ciência & Saúde Coletiva 9* (2004), 897–908.

[7] Scott, S. L. *bsts: Bayesian Structural Time Series*, 2021. R package version 0.9.7.

[8] Xavier, D. B. Estudo ecológico de séries temporais das doenças tropicais negligenciadas, malária e tuberculose-brasil, 2008 a 2030.

Métodos Estatísticos de Classificação: Abordagem Aplicada ao Diagnóstico de Casos de Câncer de Mama

Paulo Victor Cunha Porto (UFF)
Jessica Quintanilha Kubrusly (UFF)

Email de contato: pauloporto@id.uff.br, jessicakubrusly@id.uff.br.

Resumo

Este trabalho analisou o desempenho de métodos de classificação em um banco de dados de paciente com tumores malignos e benignos. Os métodos utilizados foram: Regressão Logística, Florestas Aleatórias, SVM Polinomial e SVM Radial. Todos os métodos apresentaram um bom desempenho, com níveis de acerto na base de teste acima de 90,0%. Ainda foi possível identificar o raio do tumor como a covariável de maior impacto nas chances de diagnóstico de câncer de mama.

Palavras-chave: Câncer de Mama, Regressão Logística, Floresta Aleatória, *SVM*.

Introdução

Segundo dados do Instituto Nacional do Câncer [3], o câncer de mama é um dos três tipos de maior incidência no mundo, com 2,1 milhões de casos anuais. Porém, entre as mulheres é o tipo mais frequente, representando 24,2% dos casos de câncer na população feminina. O mesmo estudo estima que, no Brasil, 66.280 casos novos de câncer de mama, para cada ano do triênio 2020-2022, valor correspondente a 61,61 casos de câncer a cada 100 mil mulheres. Seguindo o panorama global, no país o cenário é similar e o câncer de mama também é o tipo mais comum em mulheres.

Alguns dados e estudos, como por exemplo a base de dados Atlas On-line da Mortalidade[1] do Instituto Nacional do Câncer e o resumo do terceiro relatório de especialistas do Ministério da Saúde [8], relatam a relação desta doença com algumas características ou hábitos dos pacientes: idade, prática de atividades físicas regulares, boa alimentação, não fumar e não ingerir bebidas alcóolicas. Entretanto, a despeito de bons hábitos contribuirem para evitar a doença, o diagnóstico precosse continua sendo imprescindível. Para contribuir neste debate, este texto se propõe em apresentar uma relação entre o diagnóstico de câncer e características do tumor, extraídas de exames de imagens. Para isso foi considerado o banco de dados extraído da plataforma Kaggle[2], contendo informações sobre características físicas de tumores de mamas, alguns benignos outros malignos.

Este trabalho encontra-se dividido nas seguintes seções: Materiais e Métodos, onde serão apresentados os métodos aqui utilizados; e Resultados e discussão, onde encontram-se as evidências empíricas obtidas a partir da aplicações dos métodos e relatos dos principais achados do trabalho.

Material e métodos

Em geral, nos problemas de classificação, como o abordado neste texto, é considerada uma variável resposta $Y_i \sim Bernoulli(\pi_i)$ indicando certa característica de interesse e um conjunto de p covariáveis $\{X_1,X_2,\ldots,X_p\}$ em uma amostra de tamanho n. Os modelos de classificação buscam prever o valor da variável resposta Y_i a partir da observação das covariáveis $X_{i,j}$, $i = 1,2,\ldots,n$ e $j = 1,2,\ldots,p$.

Para este trabalho, em particular, as unidades amostrais são tumores de mama. A variável resposta $Y_i = 1$ se o i-ésimo tumor for classificado como maligno (positivo para a doença) e $Y_i = 0$ se ele for classificado como benigno (negativo para a doença). As covariáveis são características físicas do tumor, quantificadas através de exames de imagem, como por exemplo: Raio do tumor, perímetro, textura, entre outras.

[1] https://www.inca.gov.br/app/mortalidade
[2] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Pré-processamento dos Dados

Antes da descrição dos métodos serão apresentadas as três etapas de pré-processamento realizadas no banco de dados em questão. Tais etapas, de forma conjunta, buscam melhorar o desempenho dos métodos, a estimativas dos parâmetros e a capacidade de comparação entre eles.

Banco de Treino e Teste: A primeira etapa é a separação do banco de dados em duas partes. A primeira, chamada de banco de treino, é composta por 80% das observações e utilizada para treinar os modelos e estimar seus parâmetros. A segunda, chamada de banco de teste, é composto pelos outros 20% das observações e será usada somente na comparação dos resultados.

Análise de Multicolinearidade: A segunda etapa é a análise de multicolinearidade, onde é construída a matriz de correlação entre todas covariáveis do banco e selecionadas um subconjuntos delas composto por covariáveis com correlação menor que 0,7 entre si.

Padronização: A terceira etapa é a padronização das covariáveis. Essa é uma etapa recomendada quando as covariáveis estão em escalas muito diferentes. O processo é simplesmente trabalhar com as covariáveis transformadas da seguinte maneira:

$$\tilde{X}_{i,j} = \frac{X_{i,j} - \mu_j}{\sigma_j}. \tag{1}$$

sendo μ_j a média amostral e σ_j o desvio padrão amostral da covariável X_j. Vale a seguinte ressalva, uma vez as covariáveis padronizadas é preciso cuidado nas interpretações dos parâmetros.

Regressão Logística

O modelo de regressão logística estabelece uma relação entre a média da variável resposta e uma transformação linear das covariáveis através da função de ligação logística g, definida por:

$$g(x) = \frac{e^x}{1 + e^x}. \tag{2}$$

Assim, a Regressão Logística define a seguinte relação entre $E(Y_i) = \pi_i$ e as covariáveis:

$$E(Y_i) = \pi_i = \frac{e^{\boldsymbol{X_i'\beta}}}{1 + e^{\boldsymbol{X_i'\beta}}} \text{ , sendo } \boldsymbol{X_i'\beta} = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}\,. \tag{3}$$

O vetor de parâmeros $\boldsymbol{\beta}$ será estimado a partir do estimador de máxima verossimilhança e a seleção das covariáveis que farão parte do modelo será feita pelo Método de Eliminação Progressiva [9]. Os valores de $e^{\hat{\beta}_j}$ são chamados de razão de chance estimada e possibilitam a interpretação do quanto aumenta a chance de se observar uma classe positiva para variações na covariável X_j [5].

Para que a Regressão Logística funcione como um método de classificação é necessário escolher um valor de corte r, definido a partir da Curva ROC [2], de forma que a classe estimada será positiva ($\hat{Y} = 1$) se $\hat{\pi}_i > r$ e negativa ($\hat{Y} = 0$) caso contrário.

Árvores de Classificação

Segundo James [4], os modelos de árvores são aqueles que envolvem a estratificação ou partição do espaço das observações em regiões mais simples.

A ideia principal é criar uma medida de impureza capaz de quantifica o quanto homogênea são as classes das variáveis respostas cujas covariaveis pertencem a uma certa região $R \in \mathbb{R}^p$. Uma vez definida essa medida, é decidido, em cada etapa do processo, qual a covariável j e qual o valor v que melhor repartem as observações em duas regiões $R_1 = \{X_j < v\}$ e $R_2 = \{X_j \geq v\}$ de forma a minimizar a soma das impurezas nelas.

Para as Árvores de Classificação uma medida de impureza conhecida é o Índice de Gini [4], que tem sua versão para duas classes apresentada na Equação 4, sendo p é a proporção de classes positivas entre as observações pertencentes à região R. Note que o Índice de Gini assume valores pequenos (próximo de zero) caso a região R contenha prevalência de alguma classe, isto é, $p \approx 1$ ou $p \approx 0$, e que o valor máximo de G_R ocorre em $p = 1/2$.

$$G_R = 2p(1 - p) \tag{4}$$

Uma vez definida a medida de impureza e um critério de parada C que, em geral, é o número máximo de nós ou número mínimo de observações em por nó, o processo de criação de uma Árvore de Classificação segue os seguintes passos:

1. Seja S o banco de dados inicial, com todas as observações do banco de treino.
2. Escolha valores para j, $j = 1, 2, \ldots, p$, e $v \in \mathbb{R}$ que definem as regiões $R_1 = \{X_j < v\}$ e $R_2 = \{i : X_j \geq v\}$ e que minimizam $G_{R_1} + G_{R_2}$.
3. Defina $S_1 = \{s \in S \mid s \in R_1\}$ e $S_2 = \{s \in S \mid s \in R_2\}$.
4. Verifique se o critério de parada C foi atingido para S_1. Caso positivo o algoritmo terminada. Caso negativo, faça $S = S_1$ e retome o passo 2.
5. Verifique se o critério de parada C foi atingido para S_2. Caso positivo o algoritmo terminada. Caso negativo, faça $S = S_2$ e retome o passo 2.

Após a construção da árvore ainda é feito o processo de poda, para eliminar galhos e evitar problemas relacionados a superposição. Por fim, uma vez a árvore de classificação definida, são definidas regiões $\{R_1, \ldots, R_k\}$ que formam uma partição de $\mathbb{R}^p$. A previsão da classe de uma nova observação com vetor de covariaveis $\mathbf{X} \in R_k$ será definida pela classe prevalente na região R_k.

Floresta Aleatória

Ao construirmos o modelo de árvore de classificação a árvore gerada a partir de um banco de dados é uma estimativa da "verdadeira" árvore de classificação para os dados. Em outras palavras, a regra de decisão criada a partir do algoritmo é uma função da amostra observada. Ou seja, caso os dados fossem diferentes, outra regra de decisão seria criada. Buscando diminuir a variância do modelo e incorporar essas incerteza é definido o método de *Florestas Aleatórias*.

A ideia básica deste método é criar muitas árvores em vez de apenas uma, por isso o nome de Floresta. Vamos chamar de B o número de árvores criadas. Cada uma das B árvores será criada da seguinte maneira [4]:

1. Seja S o banco de dados inicial, com todas as observações do banco de treino.
2. Selecione uma amostra de tamanho N do conjunto S, usando uma técnica de reamostragem.
3. Selecione, de forma aleatória e sem reposição, $m \approx \sqrt{p}$ entre as p covariáveis disponíveis.
4. Para a amostra com N observações e m covariáveis, crie uma árvore de classificação utilizando o método descrito na seção anterior.

Os passos $1-4$ acima são repetidos B vezes e são criadas B árvores de classificação. A previsão para uma nova observação será definida pela classe dominante entre as classificações realizadas pelas B árvores. Para ajustar o modelo de Floresta Aleatória será usado o Programa R [10] e o pacote *randomForest* [6].

Support Vector Machine (SVM)

O SVM é uma generalização do *classificador de margem máxima*, que busca definir um hiperplano que separe os dados em 2 classes [4]. Embora o *classificador de margem máxima* seja uma forma natural de resolver o problema de classificação, nem sempre existe solução para ele. A sua generalização é o chamado *Support Vector Classifier*. Nesta definição é acomodada uma "margem de erro" para classificador descrito pelo *hiperplano de margem máxima*. Quando essa margem de erro é nula, o *Classificador de Margem Máxima* e o *Support Vector Classifier* se tornam iguais.

Embora o *Support Vector Classifier* seja um método de classificação mais geral, ele ainda é criado a partir de uma regra de decisão linear. O *SVM* é uma extensão do *Support Vector Classifier* que introduz funções não lineares das covariáveis na regra de decisão. Isto é feito utilizando uma generalização do produto interno chamada de *kernels*. Os dois *kernels* utilizados nesse trabalho encontram-se destacados a seguir:

$$\textbf{\textit{Kernel} polinomial de grau d :} \quad K(x_i, x_{i'}) = \left(1 + \sum_{i=1}^{p} x_i x_{i'}\right)^d$$

$$\textbf{\textit{Kernel} radial:} \quad K(x_i, x_{i'}) = exp\left(-\gamma \sum_{i=1}^{p} (x_i - x_{i'})^2\right).$$

Para ajustar os modelos SVM será usado o Programa R [10] e os pacotes *LiblineaR* [1] e *e1071* [7].

Comparação de Resultados

Uma vez ajustado um método de classificação é definida uma regra de classificação. Cada observação do banco de teste é classificada como positivo ou negativo. Essas observações são contabilizadas em uma das 4 entradas da Tabela 1. A partir dos valores dessa tabela é possível calcular as medidas de taxa de acerto, sensibilidade e especificidade, usadas na comparação entre os modelos.

Tabela 1: Classificação de Resultados

Teste vs Doença	Doença Presente ($Y = 1$)	Doença Ausente ($Y = 0$)
Teste Positivo ($\hat{Y} = 1$)	Verdadeiros Positivos (VP)	Falsos Positivos (FP)
Teste Negativo ($\hat{Y} = 0$)	Falso Negativo (FN)	Verdadeiro Negativo (VP)

- Taxa de Acerto $= P((\hat{Y} = 1 \cap Y = 1) \cup (\hat{Y} = 0 \cap Y = 0)) = \dfrac{\text{VP} + \text{VN}}{\text{Total}}$.
- Sensibilidade $= P(\hat{Y} = 1 | Y = 1) = \dfrac{\text{VP}}{\text{VP} + \text{FN}}$.
- Especificidade $= P(\hat{Y} = 0 | Y = 0) = \dfrac{\text{VN}}{\text{FP} + \text{VN}}$.

Resultados e discussão

Os dados utilizados neste trabalho estão disponíveis e foram extraídos da plataforma Kaggle[3]. Infelizmente, não possuímos a informação sobre a unidade de medida das mesmas. As covariáveis desse banco representam características identificadas em exames de imagem de pacientes com câncer de mama. A base de dados completa é formada por 569 tumores de mama e 30 características quantitativas sobre eles, além de uma variável qualitativa indicando se o tumor é benigno ou maligno (câncer). Das 569 observações, 357 ($62,7\%$) são benignos e 212 malignos ($37,3\%$).

Primeiro o banco foi dividido em treino e teste, sendo 80% da observações selecionadas de forma aleatória para o banco de treino e o restante para o banco de teste. Considerando as observações do banco de treino, foi realizada a análise de multicolinearidade, que resultou na eliminação de 20 das 30 covariáveis. Sendo assim, os métodos foram treinados com 10 covariáveis: Raio, Textura, Suavidade, Compactação, Simetria, Dimensão Fracrtal, Textura DP, Suavidade DP, Simetria DP e Pior simetria. Estas covariáveis passaram pelo processo de padronização descrito na Equação 1.

Antes de apresentar as medidas de comparação entre todos os métodos vejamos o modelo de Regressão Logística Ajustado. A seleção das variáveis, realizada pelo Método de Eliminação Progressiva, resultou no modelo final apresentado na Tabela 2.

Tabela 2: Modelo Logístico após Eliminação Progressiva

Covariável	Coeficiente	Desvio Padrão	p-valor	Razão de Chances
Intercepto	-1.0897	0.2760	7.87×10^{-5}	Não Aplicável
Raio	5.3906	0.7141	4.38×10^{-14}	219.33
Pior_Simetria	2.2033	0.4054	5.49×10^{-8}	9.05
Textura	1.4765	0.3130	2.39×10^{-6}	4.38
Suavidade	1.3202	0.3852	6.10×10^{-4}	3.74
Suavidade_DP	0.9159	0.4459	0.039944	2.50
Simetria_DP	-1.2827	0.3693	5.14×10^{-4}	0.28

A única variável com coeficiente estimado negativo foi a Simetria_DP. Isso significa que somente para a Simetria_DP ocorre uma diminuição (efeito negativo) na chance de observarmos um tumor maligno com o aumento unitário de seu valor. Para todas as outras variáveis do modelo ocorre um aumento (efeito positivo) na chance de observarmos um tumor maligno com o aumento unitário de seus valores. Vale o destaque para a variável Raio, que a cada aumento unitário na sua medida

[3] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

estima-se um aumento médio de $4,62\times$ na chance de se observar um tumor maligno. Esse valor foi obtido após a correção realizada na razão de chance em função da padronização realizada: $e^{\hat{\beta}_j/\sigma_j^2}$.

Tabela 3: Transformação das Razões de Chances para a Escala Original dos Dados

Covariável	$\hat{\boldsymbol{\beta}}$	$\boldsymbol{\sigma}$	**Escala**	**Razão de Chances***
Raio	5.3906	3.524	01 unidade	4.62
Textura	1.4765	4.301	01 unidade	1.41
Suavidade_DP	0.9159	0.003	0.001 unid.	1.36
Suavidade	1.3202	0.014	0.001 unid.	1.10
Pior_Simetria	2.2033	0.062	0.001 unid.	1.04
Simetria_DP	-1.2827	0.008	0.001 unid.	0.85

Razão de Chances* $= exp\left(\frac{Escala \times \hat{\beta}}{\sigma}\right)$, que é a Razão de Chance na escala original.

Veja que o modelo de Floresta Aleatória também indicou a variável Raio como de maior relevância para o diagnóstico entre tumor benigno e maligno, Figura 1. Além disso, as variáveis importantes para a Floresta Aleatória, apresentadas na mesma figura, são em geral também importantes para o Modelo Logístico, Tabela 2.

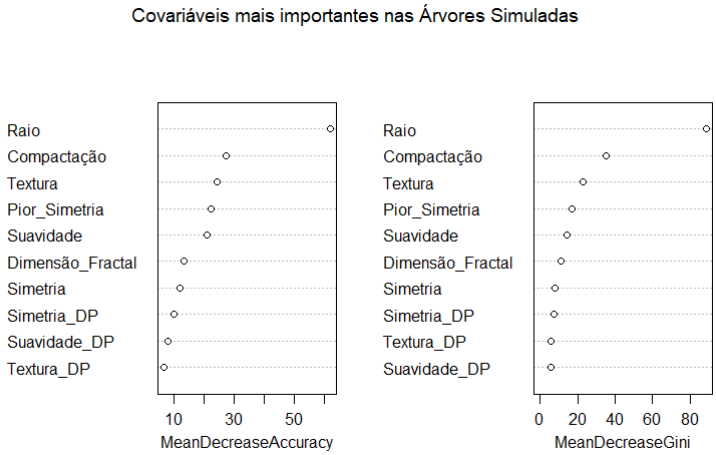


Figura 1: Importâncias das Covariáveis nas Florestas Aleatórias

A Tabela 4 apresenta uma comparação dos resultados para a amostra de teste. Todos os modelos tiveram bons resultados. Interessante destacar que o modelo de regressão logística obteve um desempenho tão bom ou melhor do que os demais modelos não paramétricos.

Tabela 4: Tabela de Resultados na Base Teste

Modelo de Classificação	Taxa de Acerto	Sensibilidade	Especificidade
Regressão Logística	**97,37%**	**100,00%**	96,25%
Florestas Aleatórias	**97,37%**	97,06%	97,50%
SVM Polinomial	**97,37%**	94,11%	**98,75%**
SVM Radial	92,98%	79,41%	**98,75%**

Tendo em vista os bons resultados, este trabalho aponta para a possibilidade de um bom desempenho de modelos estatísticos na geração de diagnósticos sobre casos de câncer de mama a partir de características físicas do tumor extraídas de exames de imagens. Essa capacidade preditiva, caso implantada tanto no setor público quanto privado de saúde, possibilitaria uma maior velocidade e escalabilidade no processo diagnóstico que hoje no país é intensivo em mão de obra médica (muitas vezes escassa no Brasil) e, por consequência, auxiliar no desfecho clínico a partir de um diagnóstico precoce da doença.

Referências

[1] Helleputte, T. *LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library*, 2021. R package version 2.10-12.

[2] Hoo, Z. H., Candlish, J. e Teare, D. What is an roc curve?, 2017.

[3] Instituto Nacional de Câncer. *Estimativa | 2020 - Incidência de Câncer no Brasil*, 2019.

[4] James, G., Witten, D., Hastie, T. e Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. Springer, 2017.

[5] Kutner, M. H., Nachtsheim, C. J., Neter, J. e Li, W. *Applied Linear Statistical Models*. McGraw-Hill, 2014.

[6] Liaw, A. e Wiener, M. Classification and regression by randomforest. *R News 2*, 3 (2002), 18–22.

[7] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. e Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2021. R package version 1.7-9.

[8] Ministério da Saúde. *Dieta, Nutrição, Atividade Física e Câncer: Uma Perspectiva Global*, 2020.

[9] Montgomery, D. C., Peck, E. A. e Vining, G. G. *Introduction to Linear Regression Analysis*. Wiley, 2012.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

Quantificando subnotificação de casos de COVID-19 no Estado do Rio de Janeiro

Ricardo Junqueira (ISP)
Jony Arrais Pinto Junior (UFF)
Rafael Erbisti (UFF)

Email de contato: rjunqueira@id.uff.br, jarrais@id.uff.br, rerbisti@id.uff.br.

Resumo

De maneira geral, avaliar as contagens de casos de uma doença, pode não fornecer a representação completa do processo de interesse. Desde o início da pandemia de COVID-19 no Brasil, as contagens da doença são marcadas por um problema de subnotificação. Neste trabalho estende-se a proposta de Stoner, Economou e Silva (2019) a partir de modificações na especificação da estrutura dos efeitos espaço-temporais para corrigir o número de casos de COVID-19 nos 92 municípios do Rio de Janeiro. Foi possível obter contagens corrigidas para cada município que são consistentes com trabalhos de outros autores que investigaram a taxa de subnotificação da doença.

Palavras-chave: processo de contagem latente. modelos CAR. COVID-19. Inferência Bayesiana.

Introdução

A COVID-19 é uma síndrome inflamatória multi-sistêmica causada pelo vírus SARS-CoV-2. Ela foi identificada pela primeira vez em novembro de 2019 na província chinesa de Wuhan e desde então espalhou-se pelo mundo. Segundo dados da Organização Mundial da Saúde, até setembro de 2021 já haviam sido confirmados 223.022.538 casos e 4.602.882 óbitos pela doença em todo o mundo. No Brasil, o primeiro caso de COVID-19 foi confirmado no Estado de São Paulo no dia 26 de fevereiro de 2020, já o primeiro óbito ocorreu cerca de um mês depois no mesmo estado, em 17 de março [5].

Diversos trabalhos recentes investigaram possíveis fatores associados a contagem de casos e de óbitos pela COVID-19. [8] apresentaram resultados que sugeriam maiores taxas de incidência e óbito na população negra em relação às demais populações no estado de Connecticut nos Estados Unidos. [13] avaliaram as características clínicas dos pacientes mortos pela doença e constataram que tanto a presença de comorbidades, como diabetes e doenças cardiovasculares, quanto idade avançada aumentavam substancialmente o risco de óbito.

De maneira geral, avaliar as contagens dos casos de COVID-19, pode não fornecer a representação completa do processo. As contagens são usualmente subnotificadas, isto é, o valor registrado é menor do que o valor verdadeiro. Nesse sentido, a incapacidade de atingir indivíduos em uma população de risco torna-se um grave problema para os gestores de saúde. No contexto estatístico, essa falta de informação é um problema que pode levar a inferências estatísticas tendenciosas, afetando as estimativas de parâmetros, previsões e incertezas associadas.

No Brasil, a forma ideal de solucionar o problema de subnotificação é melhorar o sistema de captação de informação. Em particular, para os casos de COVID-19 é ampliar a testagem para toda a população. Entretanto, esse trabalho não é trivial, pois depende de atividades do serviço local, da capacidade de gestão dos municípios e de recurso financeiro. Logo, é essencial a apresentação de métodos capazes de quantificar a incerteza das taxas de detecção de forma mais precisa, auxiliando a tomada de decisão dos gestores de saúde.

De acordo com [7] a subnotificação é, conceitualmente, uma forma de dados faltantes não intencionais, onde não é observado o número real de eventos. Uma abordagem bastante usual na literatura para se trabalhar no contexto de contagens subnotificadas foi proposta por [12]. Neste trabalho, os autores propuseram um modelo Binomial para as contagens observadas e um modelo Poisson latente para as verdadeiras contagens (não observadas), também conhecido como modelo

Pogit. Mais recentemente, [11] propuseram uma classe de modelos para correção de dados de contagem espacialmente estruturados, que utiliza um conjunto de covariáveis associado ao processo de notificação e outro conjunto de covariáveis associado ao processo de contagem. Apesar de considerarem a componente temporal em suas análises, os autores assumiram que as contagens eram independentes ao longo do tempo.

Este trabalho tem por objetivo estimar o número corrigido de casos de COVID-19 no Estado do Rio de Janeiro, usando uma generalização do modelo proposto por [11]. O modelo proposto neste trabalho modifica a maneira com que a estrutura espaço-temporal é definida. Deste modo, foi incluído um conjunto adicional de efeitos espaço-temporais, considerando efeitos temporalmente estruturados. Também foi permitido que os coeficientes associados às covariáveis variassem no tempo ou no espaço.

Material e métodos

O modelo proposto neste trabalho será aplicado no número de casos de COVID-19 nos municípios do Rio de Janeiro, utilizando como unidade de tempo as semanas epidemiológicas. Os dados referentes ao número de casos de COVID-19 e Síndrome Respiratória Aguda Grave (SRAG) foram obtidos nos portais DATASUS e SIVEP, respectivamente. É sabido que o atraso existente nas notificações torna difícil monitorar o desenvolvimento de uma doença, por isso, optou-se por utilizar um recorte temporal que vai do mês em que o Estado registrou seu primeiro caso (março de 2020) ao fim de janeiro de 2021, de modo a minimizar os efeitos deste atraso [1].

A decisão por este recorte foi reforçada pelo fato de que após este período houve a chegada de uma nova variante no Estado, que provavelmente alterou a dinâmica da pandemia e causou uma explosão no número de casos. Assim, a janela de estudo é composta pelo número de casos de COVID-19 nos 92 municípios do Rio de Janeiro, ao longo de 45 semanas epidemiológicas.

Como covariáveis associadas ao processo de contagem optou-se por utilizar as três dimensões que compõem o Índice de Vulnerabilidade Social (IVS), este índice é composto por dimensões de capital humano, renda e infraestrutura, permitindo uma boa caracterização dos municípios [4]. Para explicar a notificação foi construída a variável taxa de variação de internações por Síndrome Respiratória Aguda Grave (SRAG), sob a intuição de que taxas mais elevadas no período pandêmico quando comparado com o período pré-pandemia implicariam que mais casos de COVID-19 estariam sendo identificados. A taxa de variação de internação da SRAG foi calculada da seguinte forma:

$$TxSRAG_{t,s} = \frac{SRAG^{P}_{t,s}}{\overline{SRAG}^{AP}_{t,s}} \times 100\%, \tag{1}$$

em que $SRAG^{P}_{t,s}$ é o número de casos de internação por SRAG no município s e na semana epidemiológica t na pandemia e $\overline{SRAG}^{AP}_{t,s}$ é a média de casos de internações por SRAG nos últimos 5 anos anteriores a pandemia no município s e na semana epidemiológica t, $s = 1, \ldots, K$, $t = 1, \ldots, T$.

Para a definição do modelo, suponha uma região de interesse S formada por K áreas. Sejam $y_{t,s}$ e $z_{t,s}$, respectivamente, a contagem verdadeira, porém não observada, e a notificada (observada) na área s, $s = 1, \ldots, K$, no período de tempo t, $t = 1, \ldots, T$. Suponha também que foram observadas q covariáveis relevantes para a subnotificação $\boldsymbol{v}_{t,s} = (1, v_{1,t,s}, \ldots, v_{q,t,s})^T$ e p covariáveis $\boldsymbol{x}_{t,s} = (1, x_{1,t,s}, \ldots, x_{q,t,s})^T$ associadas com o processo de contagem do evento de interesse. O modelo é definido da seguinte forma:

$$z_{t,s}|y_{t,s}, \gamma_{t,s} \sim Binomial\left(\pi_{t,s}, y_{t,s}\right), \tag{2}$$

$$\log\left(\frac{\pi_{t,s}}{1-\pi_{t,s}}\right) = \boldsymbol{v}^T_{t,s}\boldsymbol{\beta}_t + \gamma_{t,s}, \tag{3}$$

$$y_{t,s}|\lambda_{t,s}, \phi_{t,s} \sim Poisson\left(\lambda_{t,s}\right), \tag{4}$$

$$log(\lambda_{t,s}) = \boldsymbol{x}^T_{t,s}\boldsymbol{\alpha} + \delta_t + \phi_s + \epsilon_{t,s}, \tag{5}$$

em que $\boldsymbol{\beta}_t = (\beta_{1,t}, \beta_{2,t}, \ldots, \beta_{q,t})^T$ representa o efeito das covariáveis de subnotificação na probabilidade de notificação da uma ocorrência na região s no tempo t, $\pi_{t,s}$, e $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ representa o efeito das covariáveis no processo de contagem. Além disso foi incluído um efeito aleatório não estruturado $\gamma_{t,s}$ para modelar a probabilidade de notificação, permitindo capturar o efeito de possíveis covariáveis não consideradas.

No modelo proposto, o efeito espaço-temporal considerado na Equação (5) é decomposto em 3 componentes, um efeito puramente espacial (ϕ_s) comum a todos os períodos de tempo, uma tendência temporal (δ_t) comum a todas as áreas do espaço e um conjunto de interações espaço-temporais independentes ($\epsilon_{t,s}$).

O modelo aplicado nos dados considerou $\beta_{1,t} = \beta_1$ fixo no tempo e dois parâmetros variando no tempo, o efeito aleatório temporal (δ_t) e o efeito da covariável taxa da variação da SRAG ($\beta_{2,t}$). A estrutura temporal foi definida a partir de suas distribuições *a priori* tomando como base a distribuição que define o modelo conhecido como CAR Intrínseco com a definição de uma matriz de vizinhança W_t no tempo [2, 3]. Suas distribuições são definidas a seguir:

$$\beta_{2,t}|\beta_{2,-t}, \tau_{\beta 2} \sim N\left(\frac{\sum_{j\sim i} wt_{ij}\beta_{2,j}}{\sum_{j\sim i} wt_{ij}}, \frac{1}{\tau_{\beta 2}\sum_{j\sim i} wt_{ij}}\right) \text{ e } \delta_t|\delta_{-t}, \tau_\delta \sim N\left(\frac{\sum_{j\sim i} w_{ij}\delta_j}{\sum_{j\sim i} w_{ij}}, \frac{1}{\tau_\delta \sum_{j\sim i} w_{ij}}\right), \tag{6}$$

em que wt_{ij} é o elemento da matriz de vizinhança temporal W_t que indica se os tempos i e j são vizinhos, $\tau_{\beta 2}$ e τ_δ são parâmetros de precisões associados a $\beta_{2,t}$ e δ_t, respectivamente. Foram considerados vizinhos, a semana imediadamente anterior e a posterior.

Por fim, ao conjunto de efeitos espaciais ϕ_s será atribuída a distribuição *a priori* CAR Intrínseco.

$$\phi_s|\phi_{-s}, W, \tau_\phi \sim N\left(\frac{\sum_{j\sim i} w_{ij}\phi_j}{\sum_{j\sim i} w_{ij}}, \frac{1}{\tau_\phi \sum_{j\sim i} w_{ij}}\right), \tag{7}$$

em que w_{ij} é o elemento da matriz de vizinhança W que indica se as áreas i e j são vizinhas, τ_δ é um parâmetro de precisão associado a ϕ_s. Foi adotado o critério de contiguidade para a definição de municípios vizinhos.

O modelo adotado permite que o efeito da dimensão Infraestrutura do IVS varie espacialmente com distribuição semelhante a (7). Seguindo a sugestão [11], β_1 e α_1 devem receber distribuições *a priori* informativas para garantir a convergência do modelo, $\beta_1 \sim N(-0.5, 0.6)$ e $\alpha_1 \sim N(-6, 1)$. Além disso, $\gamma_{t,s} \sim N(0, \tau_\gamma)$, $\epsilon_{t,s} \sim N(0, \tau_\epsilon)$ e todos os hiperparâmetros de precisão receberam distribuição *a priori* Gama(1, 0.5).

O ajuste dos modelos foi feito por meio de amostragem de bloco em métodos Monte Carlo via Cadeia de Markov (MCMC) com o *Automated Factor Slice Sampler*, usando o pacote *nimble* [9].

Resultados e discussão

O período de análise compreende 45 semanas epidemiológicas, iniciando na semana 13 de 2020 - semana em que foi registrado o primeiro caso no Estado - até a semana 5 de 2021. Neste período, o total de casos confirmados foi de 523.414, sendo 187.271 (35,78%) destes no município do Rio de Janeiro.

Ao longo do período observado o número médio de casos por semana epidemiológica foi de 11.632. Ao observar a série temporal nota-se também que apenas as semanas 30 (2020) e 1 (2021) registraram contagens superiores a 20.000 casos, sendo esta última o pico do período observado com 24.262 notificações. O número de casos apresentou uma forte alta no início da pandemia, sendo seguido por um período com números mais baixos entre as semanas 36 e 47 de 2020 e, então, por uma nova fase com números elevados.

Parâmetro	Estimativa	$IC_{95\%}$
α_1	-6,700	[-6,810 ; -6,568]
β_1	0,667	[0,226 ; 1,122]
C. Humano	-0,114	[-0,167 ; -0,059]
Renda	-0,399	[-0,462 ; -0,335]
τ_ϕ	0,068	[0,050 ; 0,090]
τ_δ	23,872	[14,454 ; 36,864]
τ_{SRAG}	7,830	[3,089 ; 18,972]
τ_{Infra}	7,089	[4,115 ; 13,176]

Tabela 1: Estimativas pontuais e intervalos de credibilidade de 95% das estimativas do modelo.

As estimativas pontuais e intervalares dos parâmetros do modelo são apresentadas na Tabela 1. Tanto o coeficiente da dimensão de Renda quanto o da de Capital Humano são negativos e sugerem

que quanto maior a vulnerabilidade social, menor será o número de casos de COVID-19. Entretanto, deve-se levar em consideração que cidades mais vulneráveis terão menor capacidade de testagem e, consequentemente, de identificar corretamente os casos da doença. Assim, este resultado pode estar relacionado ao fato de que as cidades com maior capacidade de testagem são aquelas com valores menores destas variáveis. Vale ressaltar que os coeficientes de infraestrutura apresentaram variação espacial. Eles foram estimados com o sinal positivo para a maioria dos municípios, indicando que municípios mais vulneráveis em infraestrutura possuem uma maior quantidade de casos.

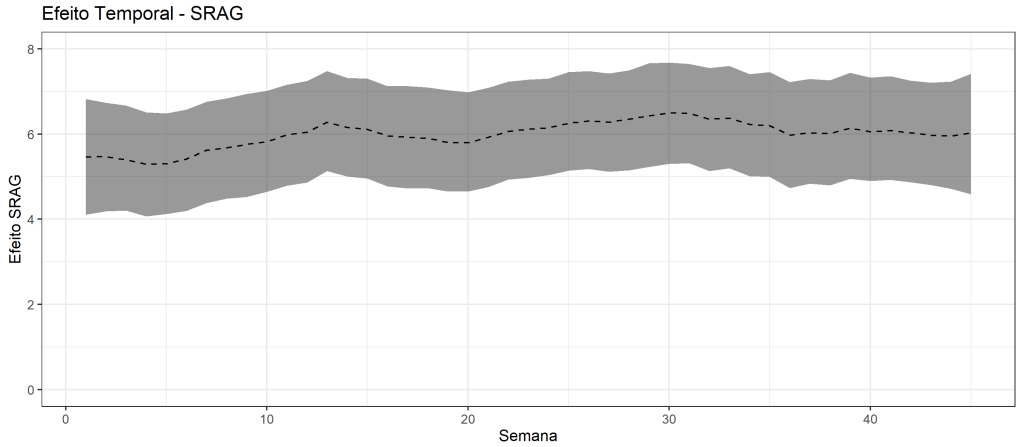


Figura 1: Estimativa pontual (linha tracejada) e intervalo de credibilidade (área cinza) do efeito da taxa de internação por SRAG ao longo das semanas epidemiológicas.

A Figura 1 apresenta a estimativa do efeito da taxa da SRAG na probabilidade de notificação. Nota-se que o efeito da taxa de internação por SRAG é positivo em todos os períodos de tempo, indicando que quanto maior for a taxa de internação, melhor é a notificação de casos. Apesar do efeito não apresentar uma grande variação temporal, o modelo apresentou uma melhor convergência com o efeito variando no tempo.

A estimativa de precisão para o conjunto de efeitos temporais sugere baixa variabilidade no número de casos ao longo do tempo. Esta precisão é substancialmente menor para os efeitos espaciais, sugerindo uma forte variabilidade espacial.

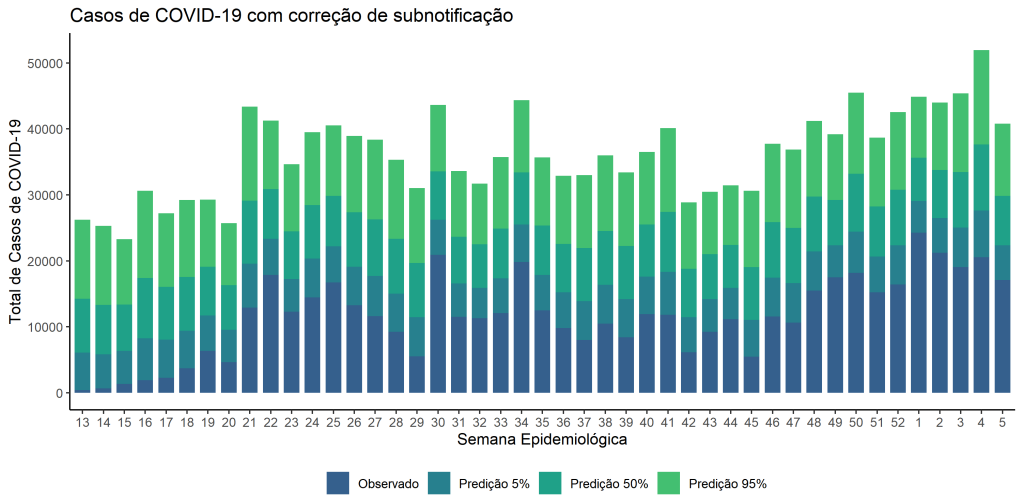


Figura 2: Série temporal do total de casos no Estado por semana epidemiológica. A linha tracejada indica o começo de 2021.

Na Figura 2, é possível visualizar, para cada semana epidemiológica, as contagens observadas da COVID-19, juntamente com os quantis de 5%, 50% e 95% da predição do número total de casos verdadeiros. Tomando como exemplo a semana 13, tem-se um total de 443 casos observados e o quantil 5% da distribuição preditiva indicaria cerca de 6.000 casos confirmados para todo o

Estado. Já o quantil 95% da distribuição preditiva indicaria cerca de 26.000 casos no Estado. Para avaliar as subnotificações para cada semana epidemiológica foram calculadas as proporções de notificação dos casos como sendo o número de casos observados/quantil 95% do número total de casos preditos. Nas primeiras 5 semanas da pandemia, a proporção de notificação dos casos esteve abaixo dos 10%. Depois de 10 semanas de pandemia este percentual se estabilizou em torno de 30% a 40% de casos notificados, atingindo um máximo de 54,06% na 43ª semana.

O modelo considerado neste trabalho incluiu dois conjuntos de efeitos espaço-temporais não estruturados, um conjunto de efeitos espaciais estruturado e um conjunto de efeitos temporais estruturado.

Os resultados obtidos neste trabalho são corroborados por resultados anteriores vistos na literatura. [6] estimaram a proporção de notificação durante o início da pandemia no Estado do Rio de Janeiro como 7,2%, [10] utilizaram dados mais completos e apontaram uma taxa de notificação de 63,6%, apontando também que regiões com melhor infraestrutura conseguem identificar corretamente mais casos. O modelo proposto tem grande flexibilidade com possibilidade de estimar probabilidades de notificação para diferentes áreas em diferentes períodos de tempo, permitindo um melhor planejamento para o enfrentamento da pandemia.

Referências

[1] Bastos, L. S., Economou, T., Gomes, M. F. C., Villela, D. A. M., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T. e Codeço, C. T. A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine 38* (2019), 4363 – 4377.

[2] Besag, J. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B. 36* (1974), 192–236.

[3] Besag, J., York, J. e Mollie, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43* (1991), 1–20.

[4] Costa, M. A., dos Santos, M. P. G., Marguti, B., Pirani, N., da Silva Pinto, C. V., Curi, R. L. C., Ribeiro, C. C. e de Albuquerque, C. G. Vulnerabilidade social no brasil: Conceitos, métodos e primeiros resultados para municípios e regiões metropolitanas brasileiras. *IPEA - Textos para discussão*, 2364 (2018).

[5] da Saúde, M. *Painel Coronavírus*, 2021.

[6] do Prado, M. F., de Paula Antunes, B. B., dos Santos Lourenço Bastos, L., Peres, I. T., de Araújo Batista da Silva, A., Dantas, L. F., Baião, F. A., Maçaira, P., Hamacher, S. e Bozza, F. A. Analysis of covid-19 under-reporting in brazil. *Revista Brasileira de Terapia Intensiva 2*, 32 (2020), 224 – 228.

[7] Gelman, A., Carlin, J., Stern, H., Dunson, D. e Rubin, A. V. D. *Bayesian Data Analysis*. Chapman and Hall/CRC Texts in Statistical Science, 2014.

[8] Laurencin, C. T. e McClinton, A. The covid-19 pandemic: A call to action to identify and address racial and ethnic disparities. journal of racial and ethnic health disparities. *Journal of Racial and Ethnic Health Disparities 7*, 3 (2020), 398 – 402.

[9] Lawson, A. B. Nimble for bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology 33* (2020).

[10] Paixão, B., Baroni, L., Pedroso, M., Salles, R., Escobar, L., de Sousa, C., de Freitas Saldanha, R., Soares, J., Coutinho, R., Porto, F. e Ogasawara, E. Estimation of covid-19 under-reporting in the braziian states through sari. *New Generation Computing* (2021).

[11] Stoner, O., Economou, T. e da Silva, G. D. M. A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association: Applications and Case Studies*, 528 (2019), 1481 – 1492.

[12] Winkelmann, R. e Zimmermann, K. F. Poisson-logistic regression. *Discussion Papers, Department of Economics, University of Munich*, 93 (1993).

[13] Zhou, F., Yu, T., Du, R. e et al. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *Lancet 359* (2020), 1054 – 1062.

Como coletar dados do *Twitter* utilizando o R

Thamires Louzada Marques (UFF)
Jessica Quintanilha Kubrusly (UFF)

E-mail de contato: thamireslouzada@id.uff.br, jessicakubrusly@id.uff.br.

Resumo

Há um grande potencial de pesquisas em dados das redes sociais, sendo o *Twitter* uma das mais famosas plataformas para exposição de opinião. O objetivo deste trabalho é apresentar um tutorial sobre como realizar a coleta de bases de dados do *Twitter* de forma automática através do Programa R. Para isso será utilizada uma API (*Application Programming Interface*) integrada com a linguagem R. Como resultado temos um banco de dados composto por *tweets* entre outras variáveis, como por exemplo, data e hora da postagem. Tanto o código utilizado quanto o banco gerado são de livre acesso pelo *GitHub* https://github.com/thamirubs/DadosTwitterSEMEST.git.

Palavras-chave: Extração de Dados, *Twitter*, R.

Introdução

O crescimento das redes sociais tem como consequência o aumento da geração de dados na internet. Os usuários são ativos e, com isso, geram dados em tempo real, de diferentes naturezas, sempre de alguma forma indicando seus interesses e opiniões. Um exemplo de rede social com usuários muito ativos é o *Twitter*, que pode ser visto como um espaço de reflexão espontânea de comportamento motivado [2].

É possível citar diferentes pesquisas com análises de dados do *Twitter*, o que indica interesse dos pesquisadores, e da população de forma geral, pela informação contida nessa rede social. O recente trabalho de Lwin et al. [4], cujo objetivo era examinar as quatro emoções (medo, raiva, tristeza e alegria) durante a pandemia, analisou mais de 20 milhões de postagens no *Twitter* durante as primeiras fases do surto da COVID-19 a partir de palavras-chaves de pesquisa. Já o trabalho de Ayo et al. [1] aplica métodos de aprendizado de máquina na classificação de discursos do ódio em textos extraídos da mesma rede social. Para esses dois estudos, e para muitos outros, foi necessário realizar uma coleta de dados no *Twitter*. E mais que isso, esses estudos só são viabilizados se a coleta é feita de forma automática e programada.

Nesse contexto, o objetivo do presente trabalho é apresentar um tutorial de como realizar uma coleta de dados da rede social *Twitter* de forma automática usando o Programa R [6]. Além de apresentar como a base de dados pode ser coletada, será apresentado uma base de dados criada a partir dos passos apresentados. Por fim, buscando seguir o modelo conceitual de reprodutibilidade [5], tanto o código criado para a extração do banco de dados quanto o banco de dados bruto gerado por esse código estão disponíveis no GitHub https://github.com/thamirubs/DadosTwitterSEMEST.git.

Vale destacar que foi observado que a maioria dos materiais sobre esse assunto estão em inglês. Então acredita-se que esse trabalho seja uma forma de propagar este conhecimento em português. Espera-se proporcionar maior autonomia aos pesquisadores, que poderão coletar bases de dados de acordo com seus temas de interesse.

Material e Métodos

Nessa seção serão apresentados as etapas e códigos no Programa R [6] para realizar uma coleta de dados automática e programada no *Twitter*.

Criação de uma API

Inicialmente, é necessário criar uma API (*Application Programming Interface*) do *Twitter* [1], utilizada para consultar os dados publicados na rede social. Nela, obtemos um identificador de acesso e uma senha, que permite realizar a conexão entre a interface de programação e o banco de dados do *Twitter*. A seguir, temos um passo-a-passo para criação de uma API a partir da plataforma de desenvolvedores do *Twitter*.

1. Primeiro cria-se uma conta no *Twitter*, que pode ser feito pelo endereço eletrônico (*link*) https://twitter.com/, clicando no botão "Inscreva-se".

2. Uma vez a conta criada, é necessário aplicar-se para uma conta de desenvolvedor. Para isso, acesse https://developer.twitter.com/, clique no botão *"Apply"* e em seguida *"Apply for a developer account"*.

3. Em seguida preencha a aplicação de acordo com os objetivos e características do pesquisador. Aqui nesse tutorial vamos seguir supondo a escolha de uma das opções *"Standard application"*, que tem algumas limitações, como apenas acesso a *tweets* dos últimos sete dias e a no máximo 18.000 observações (*tweets*) por pesquisa. Se essas limitações forem um problema para a pesquisa a ser desenvolvida, sugere-se solicitar pela *"Academic Research Application"*, não contemplada neste tutorial.

4. Aperte o botão *"Get started"* e você será guiado para responder algumas perguntas em inglês. Caso tenha dificuldade, plataformas de tradução podem ser úteis nesta etapa.

5. Submetida a aplicação, o *Twitter* enviará um *e-mail* informando se o acesso foi aceito. Em caso positivo, será possível acessar a plataforma de desenvolvedores pelo endereço eletrônico https://developer.twitter.com/.

6. Para conseguir usar a interação do Programa R [6] com o Twitter, você precisa definir valores para as seguintes variáveis: `app`, `consumer_key`, `consumer_secret`, `access_token` e `access_secret`. Esses valores estão disponibilizados na plataforma de desenvolvedores. Para obtê-los, siga os seguintes passos:
   - Acesse https://developer.twitter.com/;
   - Clique em *"Developer Portal"*;
   - Clique em *"Projects & Apps"* e em seguida clique em *"Overview"*;
   - Na caixa embaixo de *"STANDALONE APPS"*, há o nome escolhido por você. Esse será o valor da variável `app`;
   - Embaixo de *"Standalone Apps"*, você verá um ícone de uma chave. Clique na chave.
   - Vamos agora gerar os valores de `consumer_key` e `consumer_secret`: clique no botão *"Regenerate"* ao lado de *"API Key and Secret"*. Uma janela vai ser aberta com os valores de *API Key* (`consumer_key`) e *API Key Secret* (`consumer_secret`).
   - Por fim, vamos gerar os valores de `access_token` e `access_secret`: clique no botão *"Regenerate"* ao lado de *"Access Token and Secret"*. Uma janela vai ser aberta com os valores de *Access Token* (`access_token`) e *Access Token Secret* (`access_secret`).

Criada a API, passamos para os processos feitos no R, com a chave e a senha obtidas.

Coleta dos Dados no R

Nesta seção será apresentado um código (*script*), na linguagem de programação R [6], utilizando o pacote *rtweet* [3] para acessar a API e coletar os dados. Também será usado o pacote *dplyr* [8] para tratamento da base. No código apresentado é realizada uma consulta à lista dos 50 *trending topics*

[1] https://developer.twitter.com/

do *Twitter* no Brasil, isto é, tópicos que estão sendo frequentemente comentados pelos usuários nas últimas horas no Brasil. Dessa lista, são selecionados somente os *trending topics* compostos por *hashtags* (termos iniciados pelo símbolo #). São então coletados até 18.000 *tweets* em português contendo essas *hashtags*.

Veja abaixo o código que realiza esse procedimento. Logo no início do código, dentro da função `create_token`, os valores para `app`, `consumer_key`, `consumer_secret`, `access_token` e `access_secret` devem ser aqueles encontrados no passo 6 apresentado na Seção de Criação de uma API deste trabalho. Os valores apresentados são valores fictícios.

```
library(dplyr)
library(rtweet)

create_token(
   app = "XXXXXXXXXXXXXXX",
   consumer_key = "xXxXxXxxXXXxxXXxxXX",
   consumer_secret = "XX1XxXxxXXxXxx11xxxXXX1X111XX1xXxxXXxxX1X11Xxx1xXx",
   access_token = "111111111111111111-XX11XX1X1xxxXxX11xxXxx1111XXxX",
   access_secret = "XXxxXXXXXXX1xXXXXxxXxXXx1xXXx1XxXx1xx1X1xXxxx")

# pegar o codigo woeid (Where On Earth IDentifier) para o brazil
aux <- trends_available()
woeid_brazil <- aux$woeid[which(aux$name == "Brazil")]

# pegar #top50 trending topics do brazil
trends_brazil <- get_trends(woeid=woeid_brazil)

# filtrar os trending topics do brazil com hashtags
trends_brazil_com_hashtags <- trends_brazil |>
  filter(grepl("#", trend)) |>
  select(trend) |>
  pull()

# criar objeto da classe character com as hashtags separadas por OR
trends_para_consulta <- paste0(trends_brazil_com_hashtags, collapse=" OR ")

# consulta
tweets <- search_tweets(q=trends_para_consulta,
                        n=18000,
                        include_rts=FALSE,
                        lang="pt")
```

Seleção das Variáveis de Interesse e Criação de Arquivo `.RData`

Após a coleta, o objeto `tweets` criado é um banco de dados, com 90 variáveis. Nem todas as 90 variáveis são do nosso interesse e por isso foi feita uma seleção das variáveis que queremos salvar no banco. As variáveis selecionadas foram: `"created_at"`, `"screen_name"`, `"text"`, `"display_text_width"`, `"is_quote"`, que foram renomeadas e estão descritas na Tabela 1.

Tabela 1: Tabela de Variáveis

Variável	Tipo	Descrição
DATE_TIME	Caractere	data e hora de publicação
USERNAME	Caractere	nome do usuário publicante
TEXT	Caractere	texto publicado
TEXT_WIDTH	Numérico	número de caracteres do texto publicado
IS_QUOTE	Lógico	variável indicadora: 1 se a publicação é uma repostagem de terceiros e 0, caso contrário

Em seguida, para cada *hashtag* popular, é criada uma variável do tipo lógico que indica a presença ou ausência dela no *tweet* coletado. A base tratada é então salva em um diretório local.

O código para executar a seleção de variáveis e salvar a base de dados no diretório como um arquivo `.RData` é apresentado a seguir.

```
# selecionar variaveis de interesse
tweets <- tweets |>
  select(DATE_TIME = created_at,
         USERNAME = screen_name,
         TEXT = text,
         TEXT_WIDTH = display_text_width,
         IS_QUOTE = is_quote)

# criar variaveis logicas para cada hashtags
presenca_hashtags <-
   sapply(
      (1:length(trends_brazil_com_hashtags)),
      FUN = function(x){
         grepl(trends_brazil_com_hashtags[x],tweets$TEXT)
      }
      )

colnames(presenca_hashtags) <- trends_brazil_com_hashtags

tweets <- cbind(tweets, presenca_hashtags)

# salvar arquivo
filename <- paste0("TwitterData_",
                   strftime(Sys.time(), format="%d%m%Y_%H"),
                   "h.RData")

# # salvar arquivo .RData
save(tweets, file=filename)
```

Execução Programada

A execução programada foi feita no RStudio [7] através do Pacote *taskscheduleR* [9]. Primeiro instale e carregue o pacote. Em seguida, para programar a execução de um arquivo `.R` siga os seguinte passos:

1. No menu superior no RStudio clique em *Tools*.
2. Selecione *Addins* e então clique em *Browse Addins*.
3. Selecione a linha referente ao pacote *taskscheduleR* e clique em *Execute*.
4. Uma janela será aberta, e nela serão preenchidas as opções da programação de execução do código, como o arquivo `.R` que será executado; a frequência que ele será executado (diária, semanal, mensal) e a data de início.

Esse código foi programado para ser executado automaticamente todos os dias, três vezes por dia: às 10:00 da manhã, às 15:00 da tarde e às 20:00 da noite.

Resultados

Esse processo de coleta, limpeza e tratamento deu-se de 01/07/2021 a 01/10/2021 (inclusive). A base de dados gerada e o código (*scripts*) do R estão disponíveis no **GitHub** https://github.com/thamirubs/DadosTwitterSEMEST.git.

Os dados coletados estão disponibilizados na seguinte estrutura: há dois bancos de dados no formato ".RData"para cada mês. O primeiro, com sufixo "_1"refere-se a datas menores ou iguais a 15/xx/2021 e o segundo, com sufixo "_2"refere-se a datas maiores ou iguais a 16/xx/2021. Os dados coletados para o dia 01/10/2021 estão no último banco de dados do mês de setembro, veja a Tabela 2. Nesta tabela também são apresentadas a quantidade de linhas (observações) e colunas (variáveis) de cada base. A quantidade de variáveis apresentada na última coluna da Tabela 2 é referente às 5 variáveis apresentadas na Tabela 1 mais as variáveis indicadores criadas para cada *hashtags* coletada.

Tabela 2: Tabela de Descrição das Bases de Dados

Arquivo	Datas	Linhas	Colunas
TrendingTopicsTwitterBrasil202107_1.RData	01/07/2021 a 15/07/2021	492.529	220
TrendingTopicsTwitterBrasil202107_2.RData	16/07/2021 a 31/07/2021	461.238	216
TrendingTopicsTwitterBrasil202108_1.RData	01/08/2021 a 15/08/2021	338.319	208
TrendingTopicsTwitterBrasil202108_2.RData	16/08/2021 a 31/08/2021	301.434	143
TrendingTopicsTwitterBrasil202109_1.RData	01/09/2021 a 15/09/2021	350.732	173
TrendingTopicsTwitterBrasil202109_2.RData	16/09/2021 a 01/10/2021	267.100	126

Dessa forma, o código apresentado neste trabalho, ao ser rodado do dia 01/07/2021 ao dia 01/10/2021, diariamente às 10:00h, 15:00h e 20:00h, gerou uma base de dados composta por um total de 2.347.024 *tweets* associados às *hashtags* dentro da lista dos 50 *trending topics* no instante da coleta. Ao longo dos 93 dias de coleta foram identificadas 864 *hashtags* distintas na lista dos termos mais comentados. Além disso, depois de 279 coletas realizadas, foi observado um total de $8,61\%$ das publicações sendo repostagens (*retweets*) de terceiros e as demais publicações sendo autênticas (escritas pelo próprio usuário).

Referências

[1] Ayo, F. E., Folorunso, O., Ibharalu, F. T. e Osinuga, I. A. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review 38* (2020), 100311.

[2] Cram, L., Llewellyn, C., Hill, R. e Magdy, W. Uk general election 2017: A twitter analysis. *arXiv preprint arXiv:1706.02271* (2017).

[3] Kearney, M. W. rtweet: Collecting and analyzing twitter data. *Journal of Open Source Software 4*, 42 (2019), 1829. R package version 0.7.0.

[4] Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R. e Yang, Y. Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. *JMIR public health and surveillance 6*, 2 (2020), e19447.

[5] Mondelli, M. L., Peterson, A. T. e Gadelha, L. M. Exploring reproducibility and fair principles in data science using ecological niche modeling as a case study. In *International Conference on Conceptual Modeling* (2019), Springer, pp. 23–33.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[7] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.

[8] Wickham, H., François, R., Henry, L. e Müller, K. *dplyr: A Grammar of Data Manipulation*, 2021. R package version 1.0.7.

[9] Wijffels, J. e Belmans, O. *taskscheduleR: Schedule R Scripts and Processes with the Windows Task Scheduler*, 2021. R package version 1.5.

Instituições parceiras e patrocinadores

A 12ª Semana da Estatística fez parte da Agenda Acadêmica da Universidade Federal Fluminense de 2021. Ocorrendo de forma remota, neste edição, não contou com recursos financeiros desta nem de outras instituições.