enacomp

ANAIS DO

XVI ENCONTRO ANUAL DE COMPUTAÇÃO

editores

Núbia Rosa da Silva

Douglas Farias Cordeiro

SBC Sociedade Brasileira de Computação

UFCAT CIÊNCIA DA COMPUTAÇÃO

Organizadores

Núbia Rosa da Silva
Douglas Farias Cordeiro

ANAIS DO XVI ENCONTRO ANUAL DE COMPUTAÇÃO

Catalão
UFCAT
2021

COORDENAÇÃO EDITORIAL
Núbia Rosa da Silva
Douglas Farias Cordeiro

COMISSÃO ORGANIZADORA DO EVENTO
Liliane do Nascimento Vale – Coordenação
Núbia Rosa da Silva – Vice Coordenação

COMISSÃO CIENTÍFICA
Douglas Farias Cordeiro
Larissa Machado Vieira
Liliane do Nascimento Vale
Núbia Rosa da Silva
Ricardo C. A. da Rocha
Tércio Alberto dos Santos
Thiago Jabur Bittar

CAPA
Douglas Farias Cordeiro

NORMALIZAÇÃO
Douglas Farias Cordeiro

DESIGN GRÁFICO E DIAGRAMAÇÃO
Douglas Farias Cordeiro

Universidade Federal de Catalão
Reitora – Roselma Lucchese

Instituto de Biotecnologia
Diretor – Geraldo Sadoyama Leal
Vice-diretor – Marcos Aurélio Batista

2021

Endereço:
Av. Dr. Lamartine Pinto de Avelar, 1120
Setor Universitário, Catalão - GO - CEP 75704-020

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)

Encontro Anual de Computação (16.: 2021, Catalão, GO)

E56a Anais do Encontro Anual de Computação [recurso eletrônico] / organizadores Núbia Rosa da Silva, Douglas Farias Cordeiro. – Catalão : UFCAT, 2021.
p. 91

ISSN: 2178-6992

1. Ciência da computação. 2. Indústria 4.0. 3. Tecnologia. I. Da Silva, Núbia Rosa. II. Cordeiro, Douglas Farias. CDU 02+004.6.

CDU: 020

Apresentação

Seja bem-vindo ao XVI Encontro Anual de Computação (EnAComp) 2021, sediado na Universidade Federal de Catalão (UFCat), e organizado pelo Departamento de Ciência da Computação (DCC). O evento, anualmente, reúne professores, pesquisadores, profissionais e estudantes da graduação e pós-graduação de todo Brasil, com o objetivo de discutir sobre as inovações referentes à Computação.

O objetivo primário do EnAComp é despertar o interesse de alunos da computação e de áreas correlatas, pelos temas que têm estado em destaque na academia e na indústria, afim de reiterar os alunos e profissionais às necessidades da pesquisa e do mercado de trabalho.

O evento que começou, em 2003, com o nome de Simpósio Anual de Computação (SiAComp) passou a ser denominado de Encontro Anual da Computação (EnAComp), em 2010, iniciando uma nova jornada de busca por seu reconhecimento no cenário nacional e internacional.

Em 2010, em sua 8a edição, o EnAComp teve como tema "Computação, Inovação e Mercado", trazendo profissionais nacionais e internacionais para ministrarem palestras e minicursos. O evento foi reformulado sob a coordenação dos professores Dra. Luanna Lopes Lobato e Dr. Thiago Jabur Bittar, proporcionando, além das palestras e minicursos, o Campeonato de Jogos Digitais, Maratona de Programação, apresentação de artigos orais e pôsteres e premiações. Os artigos foram apresentados por pesquisadores de diferentes regiões do Brasil, com foco voltado para o mercado de trabalho, graduação, pós-graduação e tendências da computação. Além de ter contado com o apoio de parceiros que tem, desde então, acreditado no evento, como Capes, CNPq e Fapeg, bem como os patrocinadores locais que tem nos ajudado a tornar o evento mais atraente.

Em 2011 o EnAComp em sua 9a edição, teve como tema "Tecnologias Inteligentes: Desafios Científicos e Tecnológicos na Computação", sendo coordenado pelos professores Dr. Dalton Matsuo Tavares, Dra. Liliane do Nascimento Vale e Dr. Vaston Gonçalves da Costa.

Em 2013, em sua 10a edição, o EnAComp teve como tema "Computação: da teoria à prática" com 4 dias de programação, em uma edição especial, comemorando 10 anos de evento, sendo coordenado pelos professores Dra.Luanna Lopes Lobato e Dr. Thiago Jabur Bittar. Neste ano, as palestras e minicursos abordaram assuntos referentes à Computação em suas várias vertentes e como essas são aplicadas na prática das empresas.

Em 2014, 11a edição, o evento teve como tema "Sistemas Embarcados: novas visões de desenvolvimento", sendo apresentados métodos computacionais de desenvolvimento em sistemas embarcados, ferramentas de síntese de circuitos digitais e projetos em redes de comunicação. O evento foi coordenado pelos professores Dr. Tércio A. S. Filho e Dr. Sérgio Francisco da Silva.

Em 2015, em sua 12a edição, o EnAComp teve como tema: "Computação: Tecnologia, Educação e Mercado", com o objetivo de apresentar como a computação, se relacionada às tecnologias digitais, educação e mercado, pode auxiliar nas mais diversas atividades, gerando resultados satisfatórios, seja na área científica, tecnológica, mercadológica, dentre outras. O evento foi coordenado pelos professores Dra. Luanna Lopes Lobato e Dr. Márcio Antônio Duarte.

Em 2017, 13a edição, o tema foi referente a "Interdisciplinaridade: Ciência, Mercado e Tecnologia", trazendo palestras que retrataram a computação sendo aplicada a diferentes áreas de pesquisa. O evento foi coordenado pelos professores Dra. Núbia Rosa e Dr. Márcio Antônio Duarte, com o apoio de outros professores do Departamento de Ciência da Computação, que coordenaram as demais atividades do evento.

Em 2018, 14a edição, o XIV EnAComp trouxe como tema principal a "Interatividade

Homem/Máquina: mesclagem da realidade e digitalidade", apresentando palestras e minicursos de instituições, como, Aptor Software, Facebook, Instituto Nacional de Pesquisas Espaciais (INPE), Oracle, UFG, UFSCar e USP. O evento foi coordenado pelos professores Dr. Márcio Antônio Duarte e Dra. Luanna Lopes Lobato.

m 2020, na 15a edição, o XV EnAComp, foi pela primeira vez 100% online devido a pandemia por COVID-19, e caracterizado como "um desafio de reeducação dos hábitos apresentados este ano à sociedade que são surpreendentes até para realidades ambientadas na tecnologia digital". Esta edição, foi portanto, a comunicação direta com o período que estamos vivendo, destacando uma palavra de ordem: Transformação. 'Quer transformar o mundo? A influência da Computação na Indústria 4.0', trazendo palestras, minicurso e apresentação de artigos. O evento foi coordenado pelas professoras: Liliane do Nascimento Vale e Núbia Rosa Silva.

O evento, desde sua 1a edição, estava engajado em atrair para o centro oeste do país, mais especificamente para Goiás, pessoas interessadas em discutir sobre temas em destaque na computação e áreas afins, trazendo para a região importantes palestrantes, minicursos, a realização da Maratona de Programação, Campeonato de Jogos Digitais e Apresentação de artigos, no formato pôster e oral, havendo premiação para estes e relevantes debates entre os envolvidos.

Neste ano, na 16a edição, o XVI EnAComp, e a segunda edição online devido a pandemia por COVID-19, trouxe o tema "Dai-me ciência e dados" em que o mercado em computação demanda profissionais com mais do que conhecimento técnico. Neste contexto, um cientista de dados vai além, buscando desenvolver competências e habilidades relacionadas a comunicação, senso crítico, capacidade analítica e interpretativa, programação e implementação de códigos de machine learning, resolução de problemas, visão sistêmica e estratégica entre outras. Especificamente, estes profissionais têm investigado aplicativos de software que realizam tarefas repetitivas ou simples conhecidos como bots (abreviatura de robôs de software). Em particular, bots sociais e de bate-papo que interage com humanos são um tópico de pesquisa recente. Da mesma forma, os bots podem ser usados para automatizar muitas tarefas executadas por profissionais e equipes de software em seu trabalho diário. Pesquisas recentes argumentam que os bots podem economizar o tempo dos desenvolvedores e aumentar significativamente a produtividade. Portanto, o objetivo deste encontro de três dias é reunir estudantes, pesquisadores e profissionais de ciência e dados, para discutir as oportunidades e os desafios de pesquisa e mercado. O evento ocorreu sob a coordenação de Liliane do Nascimento Vale e Núbia Rosa Silva Guimarães.

A publicação dos Anais é disponibilizada em formato eletrônico, com supervisão editorial de servidores da UFCat e participação científica de pesquisadores de instituições de diferentes partes do país e do mundo. Tal publicação conta com atribuição de número de ISSN 2178-6992 e é disponibilizada eletronicamente no endereço do evento: `http://www.enacomp.com.br`.

Este livro contém os artigos aceitos para apresentação online no EnAComp 2021, os quais tratam de vários temas de pesquisa e desenvolvimento em Ciência da Computação e áreas afins, sobretudo no contexto industrial.

Assim, torna-se possível a integração profissional e cultural entre os participantes, os quais possuem em comum o interesse pelo uso da computação em suas atividades. Ainda, é importante ressaltar que o evento tem contribuído, de forma positiva, para o crescimento e divulgação da UFCat, mais especificamente para o curso de Ciência da Computação, uma vez que provê meios de incentivo aos alunos e profissionais. Além da capacitação dos estudantes e profissionais, que é também facilitada por meio da realização deste evento, deve-se ressaltar a importância da pesquisa e da inovação tecnológica em computação, como força motriz para o desenvolvimento de um país.

Nesse contexto, o XVI EnAComp teve como foco englobar assuntos relacionados à computa-

ção e a ciência de dados para promover o conhecimento referente a temas inovadores. Fizemos nosso melhor para oferecer um interessante encontro online, estimulando a troca de informações científicas e inspirando novas ideias e colaborações. Estamos felizes com sua participação e esperamos vê-lo nas próximas edições EnAComp.

Liliane do Nascimento Vale

Sumário

Visualização de dados de violência contra a mulher em Minas Gerais e Goiás usando o Orange

Dayse S. Almeida, Luis Gustavo Nonato

Instituto de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP) – São Carlos, SP - Brasil

daysesa@ufcat.edu.br, gnonato@icmc.usp.br

Abstract. *The analysis of criminal patterns is important for crime prevention as it allows government agencies to design public policies aimed at the safety of vulnerable groups to that crime nature. In this sense, visualization tools make it easier to understand these patterns. Our purpose in this paper is to perform a spatiotemporal data analysis of violence against women in MG and GO using Orange. The results reveal the municipalities with the highest number of records of violence against women in MG, the period of the year in which there is greater occurrence and the relationship with the Human Development Index. The trends of this type of crime in GO are also revealed, according to the nature of the crime.*

Resumo. *A análise de padrões criminais é importante para a prevenção do crime por permitir que os órgãos governamentais projetem políticas públicas que visem à segurança de grupos vulneráveis àquela natureza de crime. Nesse sentido, as ferramentas de visualização facilitam o entendimento desses padrões. O objetivo neste artigo é realizar uma análise de dados espaço-temporais de violência contra a mulher em MG e GO utilizando o Orange. Os resultados revelam os munícipios com maior número de registros de violência contra a mulher em MG, o período do ano no qual há maior ocorrência e a sua relação com o Índice de Desenvolvimento Humano. São reveladas também as tendências desse tipo de crime em GO, de acordo com a natureza do crime.*

1. Introdução

A criminalidade urbana normalmente envolve aspectos sociais, econômicos e de infraestrutura urbana na oportunidade do crime. A violência doméstica contra a mulher e o feminicídio, no entanto, estão pouco associados às dinâmicas mais comuns da criminalidade urbana. O feminicídio, por exemplo, é um crime de ódio no qual as mulheres são assassinadas por sua condição de gênero e, normalmente cometido por alguém próximo à vítima como resultado final de violência doméstica contínua [FBSP 2021]. O Brasil é um dos países com maiores índices de feminicídio no mundo. Além disso, é um país no qual ocorre um estupro a cada oito minutos e, onde mais se morre em decorrência da criminalização do aborto [Ribeiro, 2021]. Sendo assim, é necessário que haja garantia de proteção às mulheres e meninas e, acolhimento às mulheres em situação de violência doméstica.

A compreensão dos padrões criminais é importante na prevenção do crime. Uma maneira de se encontrar esses padrões é utilizando ferramentas de visualização, já que

essas apresentam os dados em um formato mais compreensível. Ao se exibir os dados históricos de crimes de violência contra a mulher em gráficos, é possível extrair informações que ajudem a descrever as suas características. Sendo assim, o objetivo com este trabalho é organizar visualmente dados históricos, espaciais e séries temporais, de violência contra a mulher, ocorridos entre 2018 e 2021 em Minas Gerais e Goiás, a fim de se extrair informações que caracterizem esses crimes e mostrem as mudanças ocorridas ao longo do tempo. Para isso será utilizada a ferramenta de visualização de dados, Orange.

2. Trabalhos Correlatos

A análise de crimes vem sendo amplamente tratada na literatura, em trabalhos que abordam desde estatística e ciência de dados até visualização de dados e Sistemas de Informação Geográfica (*Geografic Information Systems* - GIS). O Mapeamento de Crimes, um ramo dos GIS, tem enfoque no desenvolvimento de ferramentas para explorar e analisar o comportamento espaço-temporal dos crimes. Em se tratando de dados espaço-temporais, existem numerosos esforços na literatura dedicados à identificação de *hotspots*.

Garcia *et al.* (2019) apresentam a CrimAnalyzer, uma ferramenta analítica assistida por visualização, que identifica *hotspots* espaciais e os padrões e tipos de crime associados a eles ao longo do tempo. A ferramenta permite aos usuários realizar consultas espaciais e temporais visualmente para compreender os padrões e a dinâmica temporal dos crimes, em regiões específicas de uma cidade. Nonato, Carmo e Silva (2020) apresentam um filtro de detecção de fronteiras no contexto de grafos desenvolvido para apoiar a análise visual de dados espaço-temporais. Ambos os trabalhos utilizam dados históricos de crimes para a análise de eventos criminais.

Em [Garcia-Zanabria *et al*. 2020], os autores propõem o Mirante, um sistema de visualização de mapeamento de crime no qual os dados espaciais são modelados em granularidade fina, em escala de rua. Isso permite análises espaço-temporais em grandes regiões e também em localizações específicas de uma cidade.

Com o os objetivo de verificar a relação entre as ocorrências de crime e as características das regiões analisadas, assim como neste artigo, Kadar e Pletikosa Cvijikj (2016) utilizam dados de uma rede social baseada em localização, extraem uma série de características e verificam a correlação espacial entre essas características e o número de crimes em uma região. Em [Kadar, Brüngger e Pletikosa 2017], as autoras medem a população de uma região censitária e analisam os eventos de crime do ponto de vista demográfico.

3. Materiais e Métodos

Para investigar os dados de violência contra a mulher em Minas Gerais e Goiás, foi utilizada a ferramenta de aprendizado de máquina e visualização de dados Orange[1] versão 3.30.1. O Orange é uma ferramenta que permite analisar dados por meio da criação de *workflows*.

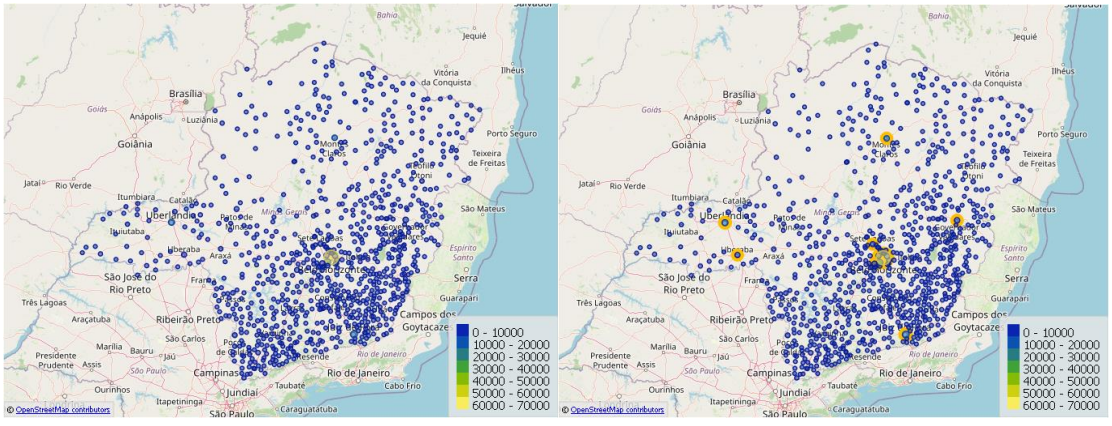
[1] https://orangedatamining.com/

Os dados abertos contendo séries temporais de violência doméstica e familiar contra a mulher e de vítimas de feminicídio nos municípios Minas Gerais têm como fonte a Polícia Civil e são disponibilizados pelo Observatório de Segurança Pública da Secretaria de Estado de Justiça e Segurança Pública (Sejusp)[2]. O período das séries temporais é de janeiro de 2018 a agosto de 2021. Os dados abertos contendo séries temporais de violência doméstica contra a mulher em Goiás têm como fonte o sistema de Registro de Atendimento Integrado (RAI) e são disponibilizados pelo Observatório de Segurança Pública da Secretaria de Segurança Pública[3]. O período dos dados analisados é de janeiro de 2018 a junho de 2021.

Adicionalmente, foram utilizados dados de 2021 de estimativas da população dos municípios[4], e da estrutura territorial[5] a fim de se encontrar a geolocalização de cada município, ambos fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Foram utilizados também dados de Índice de Desenvolvimento Humano (IDHM)[6] dos municípios, fornecidos pelo Programa da Nações Unidas pelo Desenvolvimento (PNUD) Brasil.

4. Violência Doméstica e Familiar Contra a Mulher e Feminicídio em MG

Analisando os dados de violência doméstica e familiar contra a mulher em MG, verifica-se que os 10 municípios com os maiores números de registros mensais de violência contra a mulher entre janeiro de 2018 e agosto de 2021, são: Belo Horizonte, Juiz de Fora, Betim, Contagem, Montes Claros, Sete Lagoas, Uberaba, Uberlândia, Governador Valadares e Ribeirão das Neves. Mas esses estão também entre os 11 municípios mais populosos. Na Figura 1(a) é mostrado o mapa com os municípios de Minas Gerais que possuem registros desse tipo de crime. O tamanho do ponto representa a soma do número de registros nos anos considerados, e os municípios citados estão destacados na Figura 2(b).



[2] http://www.seguranca.mg.gov.br/component/gmg/page/3118-violencia-contra-a-mulher

[3] https://www.seguranca.go.gov.br/estatisticas

[4] https://www.ibge.gov.br/estatisticas/sociais/populacao.html

[5] https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial.html

[6] https://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.html

(a) (b)

Figura 1. (a) Municípios de MG com registros de violência contra a mulher entre janeiro de 2018 e agosto de 2021. (b) Municípios com maior número de registros em destaque.

Ao analisar o espiralograma da série temporal do número de registros de violência contra a mulher em Uberlândia, mostrado na Figura 2, no qual o eixo radial se refere aos anos e o angular aos meses, observa-se que o mês com maior incidência de registros é o mês de janeiro (representado pela cor mais clara). A mesma tendência ocorre em Belo Horizonte, Juiz de Fora, Betim e Montes Claros. Governador Valadares apresenta o maior número de registros em dezembro. É possível também observar na Figura 2 que, a maior média no número de registros em Uberlândia ocorre em janeiro, mas o maior aumento no número de registros ocorre em setembro (representado pela transição abrupta de cores mais escuras para cores mais claras), seguido de agosto. A maior aceleração no número de registros ocorre em agosto. Janeiro, agosto e setembro são meses que seguem a períodos, normalmente, de férias. Juiz de Fora, Betim e Montes Claros além de possuírem a maior média, também possuem o maior aumento no número de registros em janeiro. Governador Valadares apresenta a maior média no número de registros, o maior aumento e a maior aceleração, em dezembro.

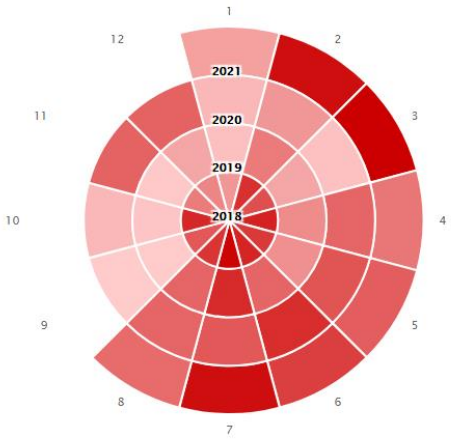


Figura 2. Espiralograma dos registros mensais de violência contra a mulher, entre 2018 e 2021, em Uberlândia, MG.

Os 10 municípios com o maior número registros de casos por mil habitantes entre 2018 e 2021 (parcial), considerando a população estimada de 2021, são: Funilândia, Água Comprida, Malacacheta, Diamantina, Ewbank da Câmara, São Gotardo, Campanário, Carmo do Paranaíba, Santana do Riacho e Couto de Magalhães de Minas. A relação entre o número de registros de casos por mil habitantes e a população estimada pode ser observada no gráfico de dispersão da Figura 3. Observa-se que Funilândia é um município com uma população pequena, 4.434 habitantes e, um número alto de registros de casos de violência contra a mulher, 47,36 registros a cada mil habitantes, em 44 meses. Funilândia é seguido por Água Comprida e Malacacheta no número de registros de casos por mil habitantes.

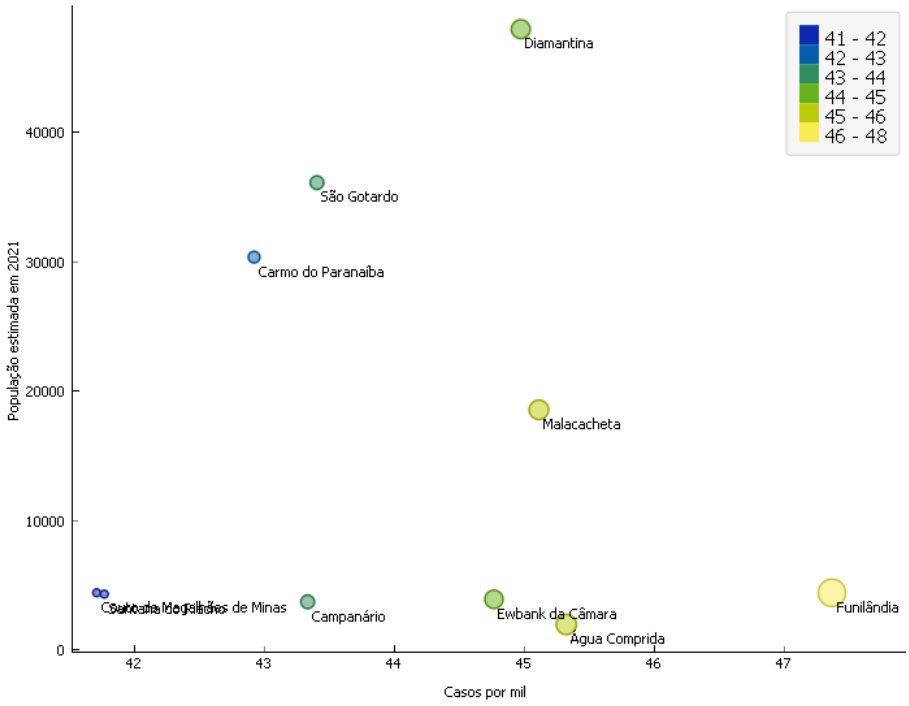
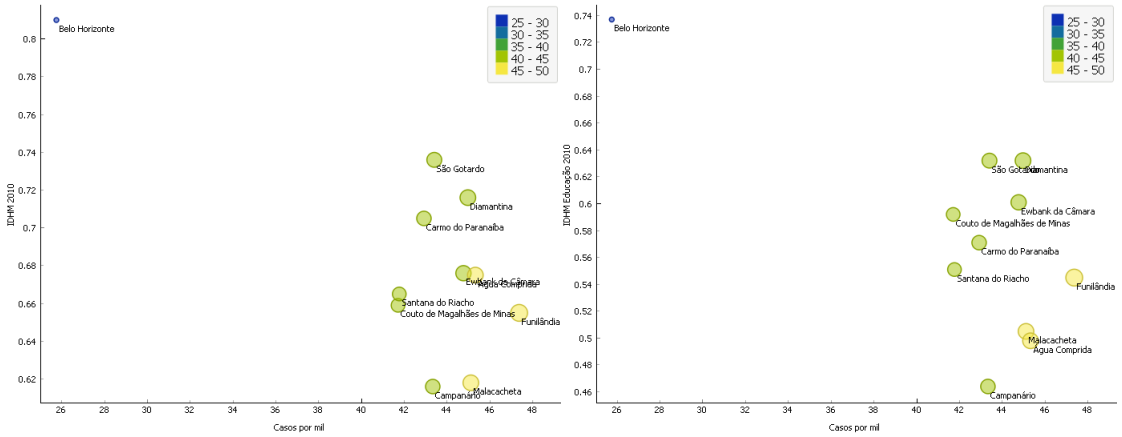
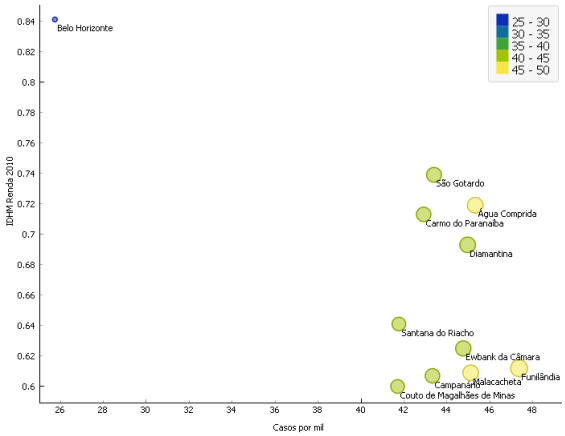


Figura 3. Os 10 municípios de MG com o maior número de registros de casos de violência contra a mulher por mil habitantes.

É possível observar no gráfico de dispersão da Figura 4(a) que o IDHM com base no Censo de 2010, é baixo nesses municípios em relação ao IDHM da capital do estado, Belo Horizonte. Enquanto que Belo Horizonte possui um IDHM igual a 0,810, Funilândia, Malacacheta e Campanário possuem IDHMs iguais a 0,655, 0,618 e 0,616, respectivamente. A razão entre o IDHM da capital e desses três municípios é ainda maior considerando apenas o IDHM em Educação, como mostrado no gráfico da Figura 4(b). Considerando apenas o IDHM em renda, a razão em relação ao índice da capital se acentua apenas para Funilândia e Malacacheta, como mostrado no gráfico da Figura 4(c).



(a) (b)



(c)

Figura 4. (a) Relação entre o número de registros de casos de violência contra a mulher por mil habitantes nos dez munícipios de MG com maior número e na capital, com o IDHM. (b) Relação com o IDHM em Educação, apenas. (c) Relação com o IDHM em renda, apenas.

Esses municípios, no entanto, possuem números de feminicídios tentados e consumados baixos, sendo que 4 dos 10 munícipios não tiveram casos nos anos considerados. Os maiores números de casos ocorreram em Diamantina, Malacacheta e Carmo do Paranaíba, com 6, 4 e 3 casos respectivamente.

Ao se considerar o feminicídio ocorrido entre janeiro de 2018 e agosto de 2021, os 10 munícipios com maior número de casos por mil habitantes são: Alvorada de Minas, Coluna, Albertina, Silvianópolis, Palmópolis, Araporã, Santo Antônio do Rio Abaixo, Datas, Matutina e Santa Fé de Minas. O IDHM nesses munícipios é inferior ao IDHM da capital do estado, como mostrado na Figura 5(a). Considerando apenas o IDHM em renda, a razão entre o IDHM da capital e dos três municípios com maiores números de feminicídio se acentua e, considerando apenas o IDHM em educação, a razão é ainda maior (Figura 5(b)).

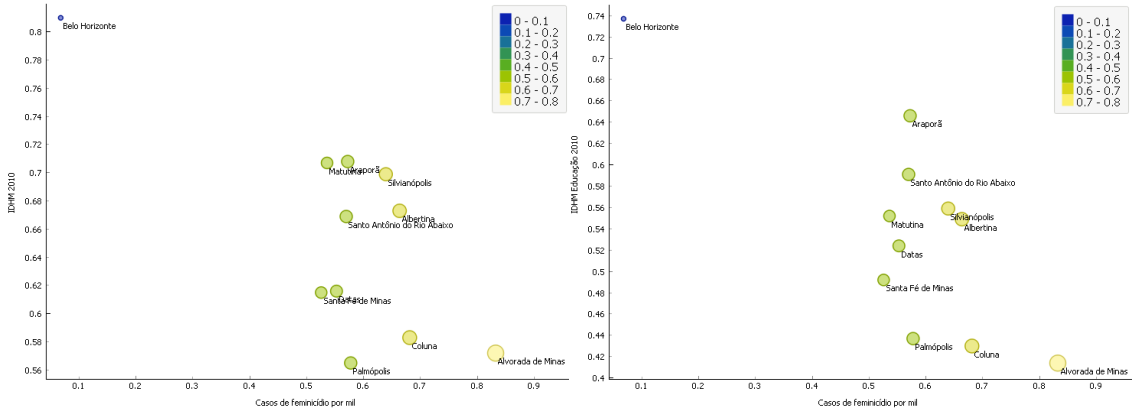


Figura 5. (a) Relação entre o número de feminicídios por mil habitantes nos dez munícipios de MG com maior número e na capital, com o IDHM. (b) Relação com o IDHM em Educação, apenas.

5. Violência Doméstica Contra a Mulher em GO

As séries temporais de registros de violência doméstica contra a mulher em Goiás são referentes apenas à natureza do crime e não aos diferentes municípios. No gráfico da Figura 6 é mostrada a evolução desse tipo de crime a partir de janeiro de 2018 até junho de 2021, por natureza do crime. Os dados de feminicídio são disponibilizados e aqui serão mostrados e discutidos em conjunto com os dados de violência doméstica. Mas é preciso observar que, de acordo com a lei 13.104/2015, femincídio é o crime praticado contra a mulher por razões da condição de sexo feminino em duas hipóteses: 1) quando o crime envolve violência doméstica e familiar; 2) quando envolve menosprezo ou discriminação à condição de mulheres. Não se sabe se os dados de feminicídio sob a hipótese 2 estão agregados aos dados de feminícidio da hipótese 1, se não houve feminicídios sob a hipótese 2 no período considerado ou, se as polícias de Goiás comumente não tipificam os homicídios femininos por menosprezo ou discriminação como feminicídio.

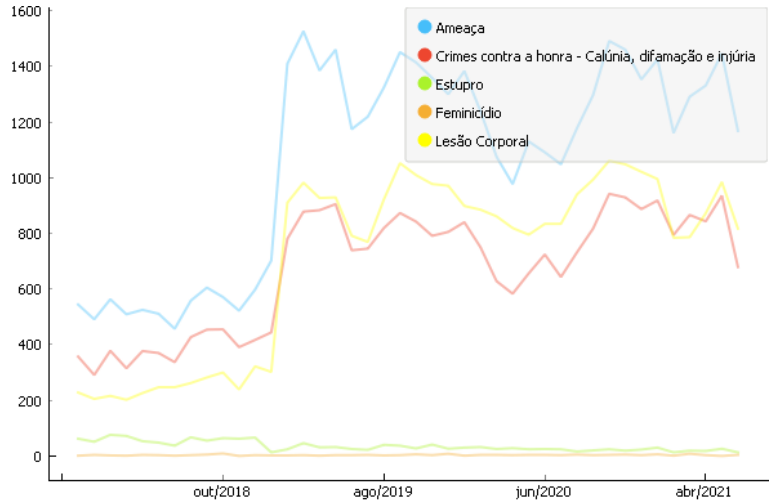


Figura 6. Números de registros de crimes de violência doméstica contra a mulher em GO entre janeiro de 2018 e junho de 2021.

Ao longo dos últimos três anos, o número anual de crime de estupro vem diminuído, enquanto que os números de crimes de lesão corporal e feminicídio vêm aumentando. Ameaça e crimes contra a honra tiveram um aumento no número de registros em 2019 e uma pequena redução em 2020, como pode ser observado no gráfico da Figura 7. É preciso considerar aqui, os impactos provocados pela pandemia de Covid-19 na vida de mulheres e meninas expostas à violência doméstica. A redução dos registros de ameaça e crimes contra a honra não significa necessariamente a redução da sua ocorrência. Observa-se no gráfico da Figura 6 que, não apenas os registros de ameaça e crimes contra a honra, como também os registros de lesão corporal, tiveram uma redução em março e, principalmente, em abril de 2020, período no qual a pandemia começava a se espalhar no Brasil, as medidas de isolamento social estavam mais rígidas e muitos serviços públicos ainda não haviam se adequado ao atendimento não presencial. Os dados agregados de 2021 não foram considerados por serem parciais.

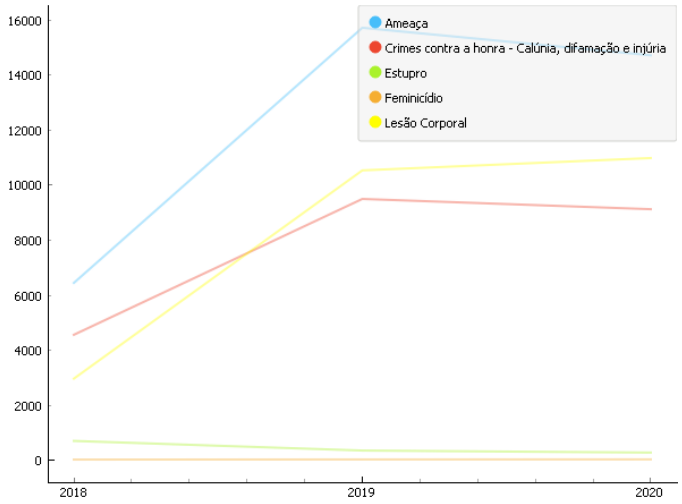


Figura 7. Números anuais de registros de crimes de violência doméstica contra a mulher em GO.

Para se analisar especificamente o aumento no número de crimes de feminicídio (Figura 8), seria necessário analisar também os números de homicídios femininos. Isso porque a lei de feminicídio é recente, de 2015, e os números apresentados dependem dos avanços que a polícia do estado de Goiás fez no sentido de investigar, entender e classificar a violência baseada em gênero. Pode ser que o aumento no número de feminicídios represente uma classificação mais adequada desse tipo de crime. No entanto, os dados de homicídios não estão disponibilizados com base no gênero.

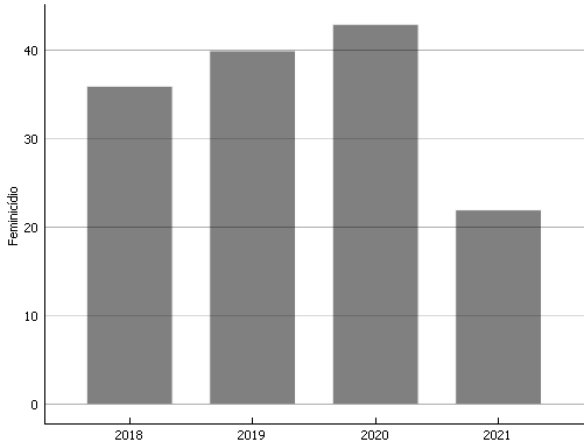


Figure 8. Número anual de feminicídios em GO

6. Conclusão

O objetivo com este artigo foi realizar uma análise visual dos dados históricos, espaciais e temporais, de violência contra a mulher em Minas Gerais e Goiás utilizando o Orange. Os resultados mostraram os municípios com maior número de registros de violência contra a mulher em Minas Gerais, o mês do ano no qual no qual ocorre maior número de registros, os munícipios com maior número de registros de violência contra a mulher e feminicído para cada mil habitantes e a sua relação com o Índice de Desenvolvimento Humano. Para o estado de Goiás, são mostradas as tendências desse tipo de crime de

acordo com a natureza do crime ao longo dos últimos três anos e, o impacto provocado pela pandemia de Covid-19 no número de registros.

Pretende-se investigar em trabalhos futuros, outros tipos de crime e a sua relação com fatores sociais, econômicos e de infraestrutura e mobilidade urbana. Para isso serão desenvolvidas técnicas baseadas na decomposição de tensores para extrair padrões de múltiplas fontes de dados, além da utilização de redes neurais recorrentes (*Recurrent Neural Network* - RNN) e de redes neurais convolucionais (*Convolutional Neural Network* - CNN) para predição de crimes em um próximo intervalo de tempo.

References

FBSP - Fórum Brasileiro de Segurança Pública (2021), Anuário Brasileiro de Segurança Pública 2021, Fórum Brasileiro de Segurança Pública, 15º edição.

Garcia, G., Silveria, J., Poco, J., Paiva, A., Nery, M. B., Silva, C. T., Nonato, L. G. (2019) "CrimAnalyzer: Understanding Crime Patterns in São Paulo", IEEE Transactions on Visualization and Computer Graphics, v. 27, n. 4, p. 1-14.

Garcia-Zanabria, G., Silveria, J., Poco, J., Nery, M., Adorno, S., Nonato, L. G. (2020). Mirante: A visualization tool for analyzing urban crimes. In *SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, p. 148-155.

Nonato, L. G.; Carmo, F. P.; Silva, C. T. (2020) "GLoG: Laplacian of Gaussian for Spatial Pattern Detection in Spatio-Temporal Data". IEEE Transactions on Visualization and Computer Graphics, v. 27, n. 8, p. 1-12.

Kadar, C., Iria, J., Pletikosa Cvijikj, I. (2016). Exploring Foursquare-derived features for crime prediction in New York City. In *International Workshop on Urban Computing (URBCOMP)*, ACM.

Kadar, C., Brüngger, R. R., Pletikosa, I. (2017) "Measuring ambient population from location-based social networks to describe urban crime". In: Social Informatics. SocInfo 2017, Editado por Ciampaglia, G., Mashhadi, A., Yasseri, T., Lecture Notes in Computer Science, p. 521–535, v. 10539, Springer, Cham, Alemanha.

Ribeiro, D. (2021) "Estupro de mulheres e feminicídio são escondidos pela imprensa patriarcal", Folha de São Paulo, https://www1.folha.uol.com.br/colunas/djamila-ribeiro/2021/10/estupro-de-mulheres-e-feminicidio-sao-escondidos-por-midia-patriarcal.shtml, Outubro.

Análise da relação entre manejos madeireiros e desmatamento usando agrupamento de séries temporais

Dayse S. Almeida, Luis Gustavo Nonato

Instituto de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP) – São Carlos, SP - Brasil

daysesa@ufcat.edu.br, gnonato@icmc.usp.br

Abstract. *In this paper we investigate, the relation between timber management and deforestation in the Amazon biome in Rondônia, using clustering of time series. These time series are related to changes in the vegetation index from 2000 to 2017 in timber management areas and in deforestation areas. The Cosine metric and the Dynamic Time Warping (DTW) algorithm were used to compute the dissimilarity between the time series, and a hierarchical clustering algorithm was used to generate the clusters. The clusters obtained in the experiments suggest a relation between timber management and deforestation in the region.*

Resumo. *Neste artigo, investiga-se a relação entre o manejo madeireiro e o desmatamento no bioma Amazônia em Rondônia, por meio do agrupamento de séries temporais. As séries temporais são relativas às alterações no índice de cobertura vegetal de 2000 a 2017 em áreas de manejo e em áreas de desmatamento. A métrica Cosine e o algoritmo Dynamic Time Warping (DTW) foram empregados para cálculo da dissimilaridade entre as séries temporais e, um algoritmo de agrupamento hierárquico foi usado para o particionamento dos clusters. As partições obtidas nos experimentos sugerem uma relação entre o manejo madeireiro e o desmatamento na região.*

1. Introdução

O bioma Amazônia, presente em nove estados brasileiros, dentre eles Rondônia, abriga uma rica biodiversidade, um quinto da água doce do planeta e tem importante papel na estabilidade do clima da América [Embrapa 2021]. O desmatamento, seja por exploração madeireira ou para a extensão das atividades agrícolas, é um dos resultados de um desenvolvimento econômico insustentável que ameaçam a Amazônia, causando a extinção de espécies de plantas e animais [Ortega *et al.* 2020]. Sendo assim, o monitoramento desse bioma é fundamental para e evitar a degradação ambiental na região.

O Instituto Nacional de Pesquisas Espaciais (INPE) desenvolve e mantêm diversos projetos para monitoramento da região da Amazônia por meio de sensoriamento remoto. As avaliações realizadas por esses projetos ajudam órgãos governamentais no controle, na prevenção e no combate ao desmatamento ilegal. Um dos projetos mantidos pelo INPE é o Projeto de Monitoramento do Desmatamento da Amazônia (PRODES) [Valeriano *et al.* 2004], [INPE 2021], que supervisiona o desmatamento em áreas com vegetação nativa desde 1988.

O Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (Ibama) realiza o controle da origem da madeira e de outros produtos florestais por meio do Sistema Nacional de Controle da Origem dos Produtos Florestais (Sinaflor), instituído em 2014 [Ibama 2021]. Sua implantação é gradual e, atualmente, apenas os estados do Pará, Mato Grosso e Minas Gerais utilizam sistemas estaduais. No entanto, se um produto florestal de origem nesses estados passa por um dos estados no qual o controle é realizado pelos órgãos responsáveis por meio do Sinaflor, ele deve possuir autorização para transporte tanto no sistema estadual do estado de origem como no sistema federal Sinaflor. Isso permite a fiscalização.

O documento que define a origem do produto florestal nativo é chamado Documento de Origem Florestal (DOF)[1], o qual permite o seu transporte legalmente. Os DOFs que possuem origem em florestas, ou seja, são emitidos para empreendimentos que realizam extração madeireira em áreas de floresta que serão chamadas aqui de áreas de manejo, devem adicionalmente estarem associados a uma AUTEX. As AUTEX se referem a autorizações ambientais emitidas pelos órgãos ambientais identificadas por meio de um número de série, além de possuírem um tipo de autorização, o ano de expedição e a localização geográfica das áreas de manejo, entre outras informações.

De acordo com a Organização das Nações Unidas para Alimentação e Agricultura (FAO) [FAO 2020a], o termo floresta se refere a uma extensão de terra maior que 0,5 hectares, coberta por árvores com mais de 5 m, cuja cobertura representa mais de 10% da superfície da área, e que é oficialmente considerada floresta. Sendo assim, qualquer evento, natural ou antrópico, que reduza permanentemente a capacidade da cobertura florestal de atingir 10% é considerado desmatamento. Ou seja, o desmatamento é a mudança permanente de uma determinada área de floresta para outro tipo de cobertura da terra [FAO 2020b]. Para o INPE, o desmatamento é um processo antrópico que pode durar vários meses [Doblas *et al.* 2020] e no projeto PRODES o desmatamento é considerado como a supressão da vegetação nativa independente de qual seja a destinação da área desmatada [INPE 2021]. Para a FAO [FAO 2020b], a remoção da cobertura florestal por manejo madeireiro não é considerada como desmatamento, pois espera-se que a floresta recupere o nível de cobertura original, de forma natural ou com o auxílio de técnicas de silvicultura.

Considerando as definições de desmatamento dadas pela FAO e pelo INPE, o objetivo neste artigo é realizar uma análise utilizando os dados do PRODES referentes a locais desmatados entre 2014 e 2017 e os dados de transporte de madeira do DOF para identificar a localização de áreas cujo manejo foi realizado entre 2014 a 2017, a fim de se verificar se existe uma relação entre os manejos madeireiros e o desmatamento no estado de Rondônia. Essa análise será realizada por meio de agrupamento de séries temporais, a partir do qual se espera obter dois grupos de séries temporais bem separados, um deles referente aos locais de desmatamento e o outro, às áreas de manejo.

[1] https://servicos.ibama.gov.br/ctf/modulos/dof/consulta_dof.php

2. Materiais e Métodos

2.1. Descrição das Bases de Dados

Foram utilizados dois conjuntos de dados para identificar as localizações das áreas desmatadas e das áreas de manejo: os dados do projeto PRODES e os dados oficiais de licenciamento de transporte de madeira dos DOFs. Para cada localização considerada, foram extraídas as séries temporais do *Portal Series View*[2] referentes às mudanças no índice de cobertura vegetal ao longo do tempo. Nesta seção, são descritos esses conjuntos de dados e apresentadas algumas análises preliminares.

2.1. 1. Dados de Desmatamento do PRODES

Os dados abertos sobre o desmatamento utilizados são aqueles organizados e disponibilizados pelo projeto PRODES no Terrabrasilis[3]. Nesse projeto é realizado o monitoramento do desmatamento por corte raso na Amazônia Legal utilizando-se imagens de sensores remotos presentes em satélites da classe Landsat, que possuem resolução espacial de 20 a 30 metros e taxa de revisita de 16 dias. As áreas desmatadas registradas pelo PRODES são de no mínimo 6,25 hectares.

No Terrabrasilis também são disponibilizados dados auxiliares sobre os limites, estados e munícipios no bioma Amazônia, além de dados do PRODES sobre áreas que não são consideradas florestas por possuírem outros tipos de vegetação.

2.1. 2. Documentos de Origem Florestal

Os transportes de produtos de origem florestal registrados no Sinaflor são legitimados por DOFs e seus dados abertos são disponibilizados pelo Ibama[4]. Esses documentos possuem a lista de produtos sendo transportados, a espécie de madeira e o volume de cada produto e, as localizações geográficas da origem e do destino. Os transportes de interesse neste trabalho se originam em áreas de manejo e, portanto, possuem uma AUTEX associada. Essas autorizações ambientais podem ser concedidas para exploração em planos de manejo florestal sustentável (PMFS), supressão de vegetação (ASV), exploração de floresta plantada (EFP) e uso alternativo do solo (UAS).

Até onde se sabe, essa é uma fonte de dados ainda muito pouco explorada.

2.1. 3. Séries Temporais

Freitas *et al.* (2011) propuseram e desenvolveram uma ferramenta para visualização de séries temporais de imagens de sensoriamento remoto, a fim de apoiar estudos de mudanças no uso e cobertura da terra. As séries temporais são derivadas de dados de imagens do sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*) a bordo dos satélites Terra (EOS-AM1) e Aqua (EOS-PM1) e representam as mudanças nos valores do índice de vegetação EVI2 (*Enhanced Vegetation Index 2*). Esse índice é

[2] http://www.dsr.inpe.br/laf/series/

[3] http://terrabrasilis.dpi.inpe.br/

[4] https://dadosabertos.ibama.gov.br/dataset/dof-transportes-de-produtos-florestais

calculado usando a refletância de superfície das bandas Vermelho (*Red*) e NIR (*near infrared*) e destaca as variações da cobertura da terra (Equação 1):

$$EVI2 = 2{,}5 * \frac{NIR - Red}{(NIR + 2{,}4 * Red + 1)} \qquad (1)$$

Como dados de sensores remotos ópticos são afetados por ruídos do próprio sensor e pela presença de nuvens cobrindo a vegetação, as séries temporais EVI2 foram filtradas usando dados auxiliares do sensor e uma Transformada *Wavelet* Discreta. Esses filtros eliminam os pixels cobertos por nuvens e, as frequências mais altas associadas a ruídos do sensor e respostas espectrais contaminadas por nuvens e sombras.

Cada dado do MODIS possui uma resolução espacial de 6,25 hectares ou 250m^2 e, assim, cada pixel possui uma área de 250m^2 e as curvas de uma série temporal representam as variações do EVI2 ao longo do tempo sobre um *pixel*. A ferramenta de visualização de séries temporais de 2000 a 2017, está disponível no Portal *Series View* já mencionado.

2.2. Espaços de Amostragem

Para realização da análise da relação entre os manejos madeireiros e os desmatamentos por meio de algoritmo de agrupamento, foram amostradas aleatoriamente localizações em dois espaços de amostragem diferentes no estado de Rondônia. O primeiro espaço de amostragem corresponde a áreas nas quais foram realizados manejos entre 2014 e 2017, com transporte registrado de madeira. Para definir essas áreas, foram excluídas as localizações em áreas não consideradas como florestas. Foram excluídas também as localizações cujas AUTEX são do tipo exploração de floresta plantada (EFP), mantendo-se assim aquelas dos tipos PMFS, ASV, UAS e aquelas com valores *null*, que representam 13,9% das localizações. Assim, foram consideradas 1.142 localizações de manejo.

Para definir o tamanho da área de manejo, foram consideradas as coordenadas geográficas de cada local de realização de manejo, definindo-se assim uma área de 250m^2, que corresponde ao tamanho do pixel do *Series View*. Foram considerados também os pixels a 250m do *pixel* central nos pontos cardeais e colateais e, os pixels a 500m do *pixel* central nos pontos cardeias, colaterais e subcolaterais. Para a obtenção das localizações dos *pixels* na vizinhança de cada localização de manejo foi utilizada a biblioteca Python Haversine 2.5.1. Assim, definiu-se uma área de manejo de 25 *pixels*, ou 1.250m^2, como mostrado na Figura 1. Na figura é destacada, no mapa de visualizações do *Series View*, as coordenadas geográficas de um manejo no ponto central (com marcador azul, na latitude -13,432889 e longitude -60,72075) e os limites de seu *pixel*, e mostrados também os *pixels* da vizinhança nas distâncias e direções definidas.

Figura 1. Exemplo de uma área de manejo 1.250m² definida a partir de uma coordenada geográfica inicial obtida do DOF. (Fonte: elaboração dos autores por meio da ferramenta *Series View*).

O segundo espaço de amostragem corresponde aos polígonos de desmatamento do PRODES. Para cada um dos 21.954 polígonos de desmatamento ocorrido entre 2014 e 2017 em Rondônia, foi calculado o centroide e utilizada a localização geográfica desse ponto e o seu *pixel* correspondente, para a definição das áreas de desmatamento.

2.3. Métodos

Para a realização dos experimentos foi utilizada a ferramenta de visualização de dados, aprendizado de máquina e mineração de dados, Orange[5] versão 3.29.3. Essa ferramenta possui a métrica de dissimilaridade *Cosine* e o método de agrupamento hierárquico. Foram utilizados também o módulo DTW do pacote Python DTAIDistance versão 2.3.2 que implementa o algoritmo *Dynamic Time Warping* (DTW), e os algoritmos de agrupamento hierárquico aglomerativo dos pacotes Python Scipy versão 1.7.1 e Scikit-learn versão 1.0. Adicionalmente, foram usados os pacotes Python Matplotlib versão 3.4.3 para gerar os gráficos de dispersão e, os pacotes Python Pandas versão 1.2.5 e Numpy versão 1.21.0 para tratar os dados.

2.3.1. Cálculo da Matriz de Dissimilaridade

A métrica *Cosine* [Tan, Steinbach e Kumar 2019], usada no Orange para calcular a matriz de dissimilaridade entre as séries temporais, é definida a partir do valor do cosseno do ângulo entre os vetores das séries temporais. Esse valor encontra-se no intervalo [-1, 1], sendo que o valor 1 representa a similaridade máxima entre duas séries temporais pois o ângulo entre os seus vetores é 0°. Para qualquer ângulo diferente de 0°, o valor de cosseno é inferior a 1. Se os vetores forem ortogonais, o valor do cosseno é 0, e se apontarem em sentido contrário, o seu valor é -1.

De acordo com Aghabozorgi, Shirkhorshidi e Wah (2015), as métricas de dissimilaridade mais comumente utilizadas para agrupamento de séries temporais são a distância Euclidiana e a DTW. Como as séries temporais utilizadas são do mesmo tamanho, ambas, *Cosine* e distância Euclidiana, se mostram adequadas. No entanto, a

[5] https://orangedatamining.com/

medida *Cosine* apresentou os melhores resultados nos testes experimentais realizados com as medidas disponíveis no Orange, dentre elas, a distância Euclidiana.

Uma desvantagem da métrica *Cosine* é que ela não é invariante ao deslocamento, ou seja, ela é calculada pontualmente com base em cada informação de tempo. Por isso, também foi utilizado nos experimentos a DTW, uma métrica de dissimilaridade cujo algoritmo encontra o mapeamento ótimo entre as séries temporais usando programação dinâmica. Assim, o objetivo do algoritmo é minimizar a distância acumulada entre duas séries temporais, sendo necessário definir uma medida de distância entre os pontos das duas séries. Esse algoritmo é muito aplicado quando o formato das séries temporais possuem variações não lineares no tempo e, assim, a similaridade entre as séries é encontrada por um alongamento ou contração não linear do eixo do tempo [Aghabozorgi, Shirkhorshidi e Wah 2015], podendo também ser aplicado em séries de tamanhos diferentes.

2.3.2. Agrupamento Hierárquico

Os algoritmos de agrupamento hierárquico [Aghabozorgi, Shirkhorshidi e Wah 2015], [Kaufman e Rousseeuw 2009] são algoritmos que constroem *clusters* aninhados dividindo-os ou fundindo-os sucessivamente. Sendo assim, eles podem ser divididos ou aglomerativos. No primeiro caso, é considerado um único grupo inicialmente, que contém todos os dados e esse grupo é iterativamente particionado até que cada instância esteja contida em grupo formado apenas por ela mesma. No segundo caso, inicialmente cada instância está contida em seu próprio grupo e, iterativamente, os grupos são fundidos para formar um novo grupo, de acordo com a similaridade entre as instâncias dos grupos, até que se tenha somente um grande grupo contendo todos os dados.

Essa hierarquia de *clusters* construída é representada como um dendrograma. A raiz do dendrograma é o aglomerado que reúne todas as instâncias e as folhas são os aglomerados que contêm apenas uma instância. O tamanho do ramo que une dois grupos representa a dissimilaridade entre eles e, pela altura dos ramos, é possível perceber visualmente o número de grupos que melhor representa a partição dos dados.

Os critérios de ligação determinam a métrica de dissimilaridade usada para a estratégia de fusão entre os grupos. O critério *single linkage* minimiza a distância entre as instâncias mais próximas de pares de *clusters*. O critério *complete linkage* minimiza a distância máxima entre as instâncias de pares de *clusters*. O critério *average linkage* minimiza a média das distâncias entre todas as instâncias de pares de *clusters*. E, finalmente, o critério *ward linkage* minimiza a variância em todos os *clusters*. O critério utilizado nos experimentos descritos a seguir foi o *complete linkage*, escolhido empiricamente.

3. Resultados e Discussões

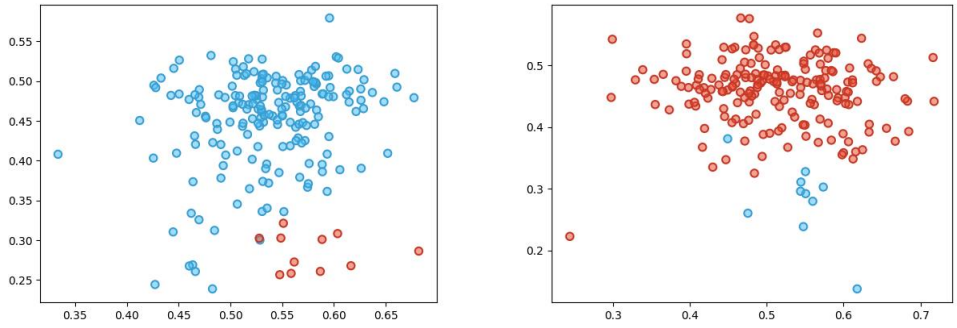
Seguindo a metodologia proposta, os experimentos se basearam em séries temporais coletadas no Portal *Series View*, filtradas por meio da Transformada *Wavelet*, e descritos a seguir.

3.1. Experimento 1 – Métrica *Cosine*

No primeiro experimento foram utilizadas 100 séries temporais de áreas de manejo de 750m^2 e 100 séries temporais de áreas de desmatamento, amostradas aleatoriamente (conjunto de dados 1). Foram utilizadas também 100 séries temporais de áreas de manejo entre 750m^2 e 1250m^2 juntamente com séries temporais de áreas de desmatamento (conjunto de dados 2). Para cada um dos dois conjuntos de séries temporais, foi calculada a matriz de dissimilaridade entre as séries usando a métrica *Cosine*, a qual alimentou o algoritmo de agrupamento hierárquico.

Nas áreas de manejo de 750m^2, com menor distância das localizações geográficas obtidas por meio dos DOFs, o algoritmo de agrupamento hierárquico gerou uma partição ruim das séries temporais. Considerando que os dados são conhecidos previamente e que o algoritmo deveria gerar dois *clusters*, a partição dos *clusters* utilizando o critério *complete linkage* foi a seguinte: o primeiro *cluster* composto por 11% das séries temporais de áreas de manejo, apenas e, o segundo, composto por 89% das séries temporais de áreas de manejo e 100% das séries temporais de áreas de desmatamento. O gráfico de dispersão baseado na correlação entre as séries temporais é mostrado na Figura 2(a). Os rótulos das séries temporais foram omitidos para facilitar a visualização. Os dendrogramas não serão mostrados devido ao grande número de séries temporais utilizado nos experimentos, o que dificulta a visualização.

Com o conjunto de dados 2, esperava-se uma dissimilaridade menor entre as séries temporais de áreas de manejo e de áreas de desmatamento, o que de fato ocorreu. Os dois *clusters* gerados foram compostos por séries temporais das duas áreas, da seguinte forma: o primeiro *cluster* composto por 91% das séries temporais de áreas de manejo e 99% das séries temporais de áreas de desmatamento e, o segundo, composto por 9% das séries temporais de áreas de manejo e 1% das séries temporais de áreas de desmatamento, como mostrado no gráfico de dispersão da Figura 2(b).



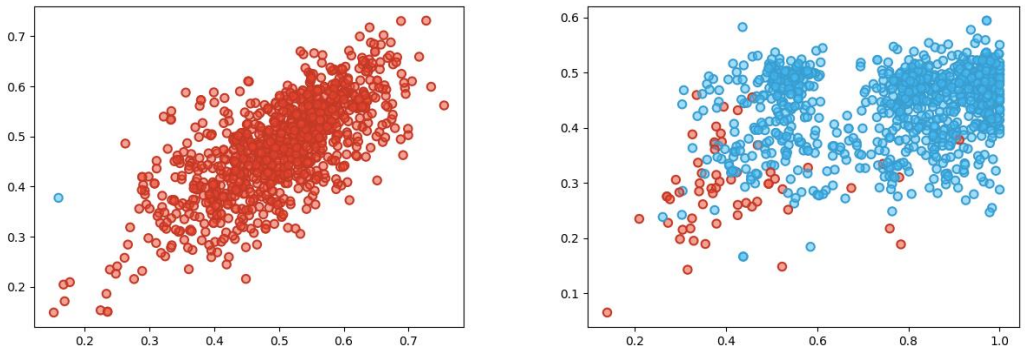
(a) **(b)**

Figure 2. (a) Dois *clusters* gerados com o conjunto de dados 1 e (b) com o conjunto de dados 2, por agrupamento hierárquico utilizando a métrica *Cosine*.

3.2. Experimento 2 – Métrica *Cosine*

Decidiu-se utilizar um número maior de séries temporais e assim, foram amostradas 500 séries temporais de áreas de manejo de 750m^2 e 500 séries temporais de áreas de desmatamento (conjunto de dados 3). Utilizando o critério *complete linkage*, o primeiro

cluster foi composto por 499 séries temporais de áreas de manejo (99,8%) e 500 séries temporais de áreas de desmatamento (100%) e o segundo *cluster* por 1 série temporal de área de manejo (0,2% das séries). Utilizando 500 séries temporais de áreas de manejo entre 750m^2 1.250m^2 e as mesmas 500 séries temporais de áreas de desmatamento (conjunto de dados 4), o primeiro *cluster* gerado foi composto por 9% das séries temporais de áreas de manejo e 2,6% das séries temporais de áreas de desmatamento e, o segundo *cluster* gerado foi composto por 91% das séries temporais de áreas de manejo e 97,4% das séries temporais de áreas de desmatamento. Os resultados com os conjuntos de dados 3 e 4 são mostrados nos gráficos de dispersão das Figuras 3(a) e 3(b), respectivamente.



(a) **(b)**

Figure 3. (a) Dois *clusters* gerados com o conjunto de dados 3 e (b) com o conjunto de dados 4, por agrupamento hierárquico utilizando a métrica *Cosine*.

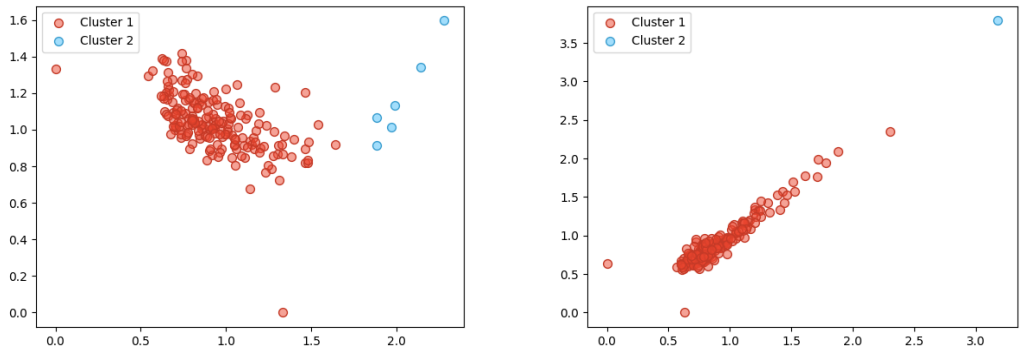
3.3. Experimentos 3 e 4 – Algoritmo DTW

No terceiro experimento foram utilizadas os conjuntos de dados 1 e 2 e, no quarto experimento, os conjuntos de dados 3 e 4. Para cada um dos conjuntos de séries temporais, foi calculada a matriz de dissimilaridade entre as séries usando o algoritmo DTW, a qual alimentou o algoritmo de agrupamento hierárquico.

Com o conjunto de dados 1, obteve-se a seguinte partição utilizando o critério *complete linkage*: o primeiro *cluster* composto por 92% das séries temporais de áreas de manejo e 100% das séries temporais de áreas de desmatamento e, o segundo, composto por 8% das séries temporais de áreas de manejo, apenas. O resultado foi similar àquele obtido com a métrica *Cosine* (Seção 3.1). O conjunto de dados 2 foi particionado da seguinte maneira: o primeiro *cluster* composto por 1% das séries temporais de áreas de manejo, apenas, e o segundo *cluster* composto por 99% das séries temporais de áreas de manejo e 100% das séries temporais de áreas de desmatamento. Os resultados com os conjuntos de dados 1 e 2 são mostrados nas Figuras 4(a) e 4(b), respectivamente.

O resultado obtido com o conjunto de dados 3 foi exatamente o mesmo que aquele obtido com o conjunto de dados 2 e, similar àquele obtido com a métrica *Cosine* para o mesmo conjunto de dados. Com o conjunto de dados 4, o resultado também foi similar ao obtido com a métrica *Cosine* para o mesmo conjunto de dados, com *clusters* desbalanceados e sem diferenciação entre as séries temporais dos dois espaços de amostragem. Assim, a partição obtida foi a seguinte: o primeiro *cluster* composto por

2,6% das séries temporais de áreas de manejo e 0,2% das séries temporais de áreas de desmatamento e, o segundo *cluster* composto por 97,4% das séries temporais de áreas de manejo e 99,8% das séries temporais de áreas de desmatamento.



(a) **(b)**

Figure 4. (a) Dois *clusters* gerados com o conjunto de dados 1 (b) e com o conjunto de dados 2, por agrupamento hierárquico utilizando o algoritmo DTW.

7. Conclusão e Trabalhos Futuros

A análise reportada neste artigo tinha como objetivo investigar a relação entre o manejo madeireiro realizado entre 2014 e 2017 e o desmatamento nesse mesmo período, nas áreas de floresta nativa de Rondônia. Essa análise foi realizada por meio de agrupamento de séries temporais que se referem às mudanças no índice de vegetação EVI2 ocorridas ao longo do tempo, no período de 2000 a 2017.

Foi definido como manejo qualquer extração de madeira da floresta nativa e, a localização geográfica foi identificada por meio dos dados de transporte de madeira dos DOFs, uma fonte de dados, até onde se sabe, ainda não explorada. Considerou-se como desmatamento a alteração definitiva da cobertura do solo em área de floresta nativa por corte raso. E, para identificar as áreas de desmatamento, foram usados dados de sensoriamento remoto, processados e disponibilizados pelo INPE.

Os resultados gerados pelo algoritmo de agrupamento hierárquico usando tanto a métrica *Cosine* quanto o algoritmo DTW, mais comumente utilizado para medir a dissimilaridade entre séries temporais, são um indício da relação entre o manejo madeireiro e o desmatamento na região. Isso porque, não foi possível separar as séries temporais dos dois espaços de amostragem em dois *clusters* distintos, contrariando a hipótese inicial. Os resultados revelaram uma forte correlação entre todos os dados. No entanto, pretende-se investigar em trabalhos futuros, as mudanças na cobertura da terra por meio das redes neurais recorrecorrentes *Long Short Term Memory* (LSTM) e mecanismos de aprendizado profundo que seguirão ao uso de codificadores automáticos (*autoencoders*) de redes neurais convolucionais. Esses codificadores serão usados para incorporar, em um espaço de alta dimensão, as séries temporais associadas a cada *pixel* da discretização espacial. Até onde se tem conhecimento, os esquemas de incorporação baseados em aprendizado profundo nunca foram usados no contexto da análise de mudanças no uso e cobertura da terra.

Referências

Aghabozorgi, S., Shirkhorshidi, A. S., Wah, T. Y. (2015) "Time-series clustering – A decade review", Information Systems, v. 53, p. 16–38.

Doblas, J., Shimabukuro, Y., Sant'Anna, S., Carneiro, A., Aragão, L., Almeida, C. (2020) "Optimizing Near Real-Time Detection of Deforestation on Tropical Rainforests Using Sentinel-1 Data", Remote Sens., v. 12, n. 23, 3922.

Embrapa (2021) "Contando ciência na web", https://www.embrapa.br/contando-ciencia/bioma-amazonia, Setembro.

FAO (2020a) "Global Forest Resources Assessment 2020: Terms and Definition", Food and Agriculture Organization of the United nations (FAO), Rome, Italy, 32 p.

FAO (2020b) "Global Forest Resources Assessment 2020 - Key Findings", Food and Agriculture Organization of the United nations (FAO), Rome, Italy, 16 p.

Freitas, R. M., Arai, E., Adami, M., Ferreira, A. S., Sato, F. Y., Shimabukuro, Y. E., Rosa, R. R., Anderson, L. O., Rudorff, B. F. T. (2011) "Virtual laboratory of remote sensing time series: visualization of MODIS EVI2 data set over South America", Journal of Computational Interdisciplinary Sciences, v. 2, p. 57-68.

IBAMA (2021) "Sistema Nacional de Controle da Origem dos Produtos Florestais (Sinaflor)", http://www.ibama.gov.br/sinaflor, Setembro.

INPE (2021) "PRODES – Amazônia", http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes, Setembro.

Kaufman, L., Rousseeuw, P. J. (2009), Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 99th Edition.

Ortega, M. X., Bermudez, J. D., Happ, P. N., Gomes, A., Feitosa, R. Q. (2019) Evaluation of deep learning techniques for deforastation detection in the amazon forest, In *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-2/W7*, p. 121–128, Copernicus Publications.

Tan, P.-N., Steinbach, M., Kumar, V. (2019), Introduction to Data Mining, Pearson Education Limited, Global Edition.

Valeriano, D. M., Mello, E. M. K., Moreira, J. C., Shimabukuro, Y. E., Duarte, V., Souza, I. M., dos Santos, J. R., Barbosa, C. C. F., de Souza, R. C. M. (2004). Monitoring tropical forest from space: the PRODES digital project. In *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, p. 272–274, ISPRS.

IPET: aplicativo mobile que conecta ongs protetoras de animais domésticos com apoiadores

Roberto Murilo M Cordeiro[1], **Valtenis R de Souza Filho**[1], **Júnio César de Lima**[1]

Instituto Federal Goiano – Campus Urutaí
Rod. Geraldo Silva Nascimento, Km-2,5 - Zona Rural, Urutaí - GO, 75790-000

{roberto.martins,valtenis.souza}@estudante.ifgoiano.edu.br

junio.lima@ifgoiano.edu.br

Abstract. *The number of abandoned animals in Brazil is growing, studies show that there are about 30 million abandoned animals, and more than 170,000 animals are under the care of NGOs protecting animals. Given these facts, this work aims to present an application built for smartphones that connects NGOs with financial supporters. The application called IPet helps NGOs in publishing cases of animals in which financial support is needed. A case has photos and a detailed description of the animal's problem, and supporters who are sensitive to a case can make a donation of any amount. The app is available for download from the Android store.*

Resumo. *Cresce o numero de animais em situação de abandono no Brasil, estudos apontam que existem cerca de 30 milhões de animais abandonados, e mais de 170 mil animais estão sob os cuidados de ONGs protetoras de animais. Diante destes fatos, este trabalho possui como objetivo apresentar um aplicativo construído para smartphones que conecta ONGs com apoiadores financeiros. O aplicativo denominado como IPet auxilia as ONGs na publicação de casos de animais em que é necessário um apoio financeiro. Um caso possui fotos e uma descrição detalhada do problema do animal, e apoiadores que se sensibilizarem com algum caso, podem fazer uma doação de qualquer valor. O aplicativo se encontra disponível para download na loja do Android.*

1. Introdução

O número de animais de estimação no Brasil está em crescente ascensão. Hoje há cerca de 54.2 milhões de cães, 39.8 milhões de aves, 23.9 milhões de gatos, 19.1 milhões de peixes e 2.3 milhões de répteis e pequenos mamíferos [Instituto Pet Brasil 2019]. Esses números mostram que o interesse das pessoas por animais de estimação está cada vez maior. Apesar do aumento da procura pelos animais, há um grande problema relacionado ao abandono de animais, uma vez o Brasil possui cerca de 30 milhões animais abandonados [Lemos 2021].

Para tentar auxiliar animais abandonados, existem as Organizações Não Governamentais (ONGs) que são entidades sem fins lucrativos que realizam ações solidárias para públicos específicos atuando em diferentes áreas [Sebrae 2017]. As maioria das ONGs voltadas para animais visam proteger animais abandonados, cuidando e aplicando os tratamentos necessários para que fiquem saudáveis e assim posteriormente doados para

famílias aprovadas pela própria ONG [Tubaldini 2019]. Segundo Velasco (2019) existem mais de 170 mil animais abandonados sob cuidado de ONGs no Brasil. Com o início da pandemia do Coronavírus (COVID-19) em 2020, este número está cada vez pior, pois feiras de doações foram proibidas e o número de adoções caíram, afetando também a doações de alimentos [TV TEM 2021].

Diante destes fatos, fica claro a importância das ONGs de animais para auxiliá-los, porém as mesmas estão cada vez mais necessitando de ajuda para se manterem. Diante deste problema, foi realizada uma pesquisa exploratória a fim de encontrar soluções que consigam conectar pessoas e ONGs de animais, onde foi encontrado diversos sites de ONGs, além de aplicações para celulares.

Um exemplo é o aplicativo **MeAuDote**, que é uma plataforma digital, voltado para doação e adoção responsável de animais que precisam de uma família, unindo de um lado as ONGs, protetores, pessoas com animais em lar temporário ou entidade que precise doar um animal e do outro lado, pessoas interessadas em adotar um animal [MEAUDOTE 2021].

Com proposta parecida, o aplicativo **Pet Ponto** funciona como um ponto de encontro para pets e potenciais tutores. Nesse aplicativo as ONGs, abrigos e tutores cadastram os pets sob seus cuidados, sendo que os interessados podem aplicar filtros por localização para encontrar os animais mais próximos [PetPonto 2021]. Já o aplicativo **Adota Pet Go** surgiu pela dificuldade na doação de animais de rua, onde ele conecta entidades e pessoas que se interessam na doação e adoção de animais, tudo de forma gratuita [AdotaPet 2021].

Ao verificar esses principais aplicativos disponíveis na loja de aplicativos do Android (*Play Store*), foi observado que a maioria deles oferecem a funcionalidade de adoção de animais, o que já possibilita uma excelente contribuição, mas como as ONGs são entidades sem fins lucrativos, e geralmente existem uma grande quantidade de animais dependendo dos seus cuidados, a parte de arrecadação de recursos financeiros é um ponto que também há a grande necessidade da participação de apoiadores.

Diante do exposto, este trabalho tem como objetivo apresentar o aplicativo para dispositivo móvel denominado IPet, que foi construído com o intuito de conectar ONGs protetoras de animais com apoiadores financeiros, além de facilitar a adoção dos animais. O IPet permite que ONGs previamente cadastradas possam incluir casos em que necessitam de apoio financeiro, informando a espécie do animal, nome, uma descrição sobre o problema do animal que precisa passar por algum procedimento em que é necessário o apoio financeiro. O valor total necessário para custear a solução do caso, também é informado pela ONG no momento do inclusão.

Já o usuário apoiador, ao acessar o aplicativo, tem acesso a lista de todos os casos, podendo aplicar filtros por ONG, região, faixa de valor e espécie do animal. Ao selecionar um caso, o apoiador tem acesso aos dados bancários ou chave Pix para realizar a doação, que pode ser de qualquer valor. Após a ONG confirmar o recebimento da doação, o valor já arrecadado do caso é atualizado.

O aplicativo IPet foi desenvolvido inicialmente apenas para *smartphones* com sistema operacional Android, no desenvolvimento foi utilizada a linguagem Java, Cloud Firestore que é um banco de dados NoSQL (*Not Only SQL*) que faz parte da plataforma de desenvolvimento de aplicativos Firebase, além do software Figma para prototipar e

validar as telas antes da implementação em código. O IPet já se encontra disponível para download na *Play Store*.

A sequência deste artigo está organizado da seguinte forma: a Seção 2 apresenta o referencial teórico necessários para desenvolvimento do aplicativo. A Seção 3 mostra os métodos utilizados desde a prototipação até o desenvolvimento. A Seção 4 aborda os resultados atingidos e discussões. Por fim as conclusões do aplicativo são mostradas na Seção 5.

2. Referencial Teórico

Essa seção realiza uma apresentação teórica e conceitual referentes às principais tecnologias utilizadas durante o desenvolvimento deste projeto, sendo elas: plataforma Android, e banco de dados NoSQL.

2.1. Plataforma Android

O Android é uma plataforma completa para dispositivos móveis que envolve um pacote com programas para celulares incluindo sistema operacional, middleware, aplicativos e interface do usuário, onde sua construção teve como objetivo permitir aos desenvolvedores criar aplicações móveis que possam tirar total proveito do que um aparelho portátil possa oferecer [Pereira and da Silva 2009].

Os aplicativos construídos para essa plataforma podem ser desenvolvidos utilizando a linguagem Java ou Kotlin, sendo o Java a linguagem adotada desde o inicio do Android. A linguagem de programação Java foi desenvolvida em 1992 pela antiga Sun por um time de desenvolvedores liderados por James Gosling, conhecido como o pai do Java, em busca de inovações tecnológicas e que "o objetivo-chave do Java é ser capaz de escrever programas a serem executados em uma grande variedade de sistemas computacionais e dispositivos controlados por computador" [Deitel et al. 2008].

O Android utiliza o arquivo $AndroidManifest.xml$ para salvar configurações de uma applicação, sendo que o arquivo contém informações como nome de classes, permissões, eventos, por exemplo. Na construção de telas, é utilizado as $atividades$ que são representadas por um arquivo XML (*eXtensible Markup Language*) para definição da interface e outro arquivo para controlar elementos exibidos ao usuários, utilizando a linguagem de programação escolhida.

O Android possui como principal vantagem, ao se comparar com seu principal concorrente iOS, ser um projeto *open-source*, onde diversas empresas disponibilizam aos seus usuários versões modificadas contendo diversas funcionalidades. Sendo assim, outra vantagem do Android em comparação ao iOS está na fácil personalização do sistema e a obtenção de novidades para *smartphones* de forma mais rápida. Segundo [Statcounter 2021] site que realiza análises de tráfego, o Android conta com uma dominância de 86.32% no Brasil se comparado as demais plataformas.

2.2. Banco de dados NoSQL

Os bancos de dados NoSQL são bancos de dados não relacionais, possuindo foco na alta performance, acessibilidade, confiabilidade e escalabilidade para uma grande quantidade de dados, além de eles conseguem armazenar dados não estruturados como *e-mail*, multimídia e documentos [Kumar and Garg 2017]. Estes bancos de dados são geralmente

classificados em orientados a colunas, chave-valor, gráficos e orientados a documentos [Gomez et al. 2016].

Um exemplo de banco de dados NoSQL é o *Cloud Firestore* sendo ele orientado a documentos que permite armazenar, sincronizar e consultar dados facilmente em apps para dispositivos móveis e da Web, além de ser possível utiliza-lo *offline* [Firestore 2021].

3. Metodologia de desenvolvimento

Na Figura 1 é mostrado um fluxograma contendo as etapas utilizadas na metodologia para o desenvolvimento do aplicativo.

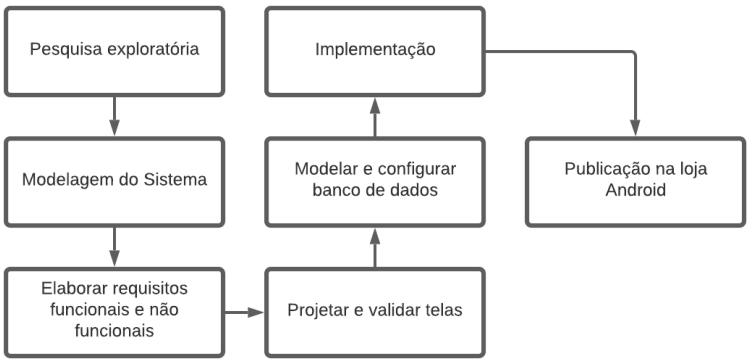


Figura 1. Fluxograma das etapas realizadas na metodologia.

Para a definição das principais funcionalidades que são fornecidas pelo aplicativo, foi realizada uma pesquisa exploratória a fim de encontrar aplicativos que possuem funcionalidades que, de alguma forma, apoiam as ONGs em seus processos de captação de recursos e facilitam a adoção dos animais amparados por elas. O principal objetivo desta etapa, foi identificar as funcionalidades que esses aplicativos não oferecem, e também, verificar recursos que já são oferecidos e que possam ser reimplementados com melhorias.

Com as funcionalidades definidas, foi realizada a modelagem do sistema, onde foram definidos os requisitos funcionais e não funcionais. Os requisitos funcionais são as funcionalidades que deverão obrigatoriamente estarem presentes na aplicação, descrevendo o comportamento que a aplicação deverá realizar. Na Tabela 1 é possível visualizar todos os requisitos funcionais que o aplicativo deverá possuir.

Ref.	**Nome**	**Descrição**
RF01	Cadastrar ONG	Possibilitar o cadastro de ONGs para gerenciar casos de animais.
RF02	Realizar login	Possibilitar o login de ONGs por meio de *e-mail* e senha.
RF03	Cadastrar Dados Bancários	Possibilitar que a ONG cadastre dados bancários para receber doações.

RF04	Criar caso de animal	Possibilitar que a ONG crie casos de animais inserindo informações e fotos.
RF05	Listar casos	Possibilitar que apoiadores e ONGs consigam visualizar a listagem de casos de animais existentes.
RF06	Filtrar casos	Possibilitar que o apoiador filtre casos baseado em nome de ONG, valor, localização e espécie.
RF07	Contatar ONGs	Possibilitar que o apoiador entre em contato com a ONG para obter informações que não estejam presentes no aplicativo.
RF08	Realizar doação	Possibilitar que o apoiador realize doações para um caso por meio de dados bancários disponibilizados pela ONG.
RF09	Gerenciar doações	Possibilitar que a ONG gerencie as doações recebidas de apoiadores.

Tabela 1. Requisitos funcionais do aplicativo

Por outro lado, os requisitos não funcionais de uma aplicação se diz a respeito aos requisitos que abordam a arquitetura, plataforma, desempenho, disponibilidade entre outros. Com isso, foram criados alguns requisitos não funcionais, sendo estes contidos na Tabela 2.

Ref.	**Nome**	**Descrição**
RNF01	Plataforma	O aplicativo deverá ser desenvolvido para plataforma Android.
RNF02	Linguagem de Programação	O aplicativo deverá ser desenvolvido na linguagem de programação *Java*.
RNF03	Arquitetura	O aplicativo deverá ser desenvolvido com uma arquitetura MVC (*Model-View-Controller*).
RNF04	Tempo de resposta	O tempo de resposta máximo permitido para transações *online* é de 15 segundos.
RNF05	Disponibilidade	O aplicativo deverá estar disponível 24 horas por dia.

Tabela 2. Requisitos não funcionais do aplicativo

Na sequência foram, projetadas e validadas as telas. Para a criação dos protótipos de tela foi utilizado o software Figma. O Figma é uma ferramenta de UI (*User Interface*) online e gratuita, feita para criar, colaborar, prototipar e inspecionar interfaces de diversos tipos.

Depois da definição das telas principais, foi realizada a modelagem e configuração da base de dados. Após analisar as opções disponíveis, optou-se pelo uso do *Cloud Fires-*

tore por ser um banco de dados NoSQL focado na integração com dispositivos móveis.

Em seguida, foi iniciada a implementação do aplicativo na plataforma Android com a linguagem Java, onde foi utilizado o Kit de Desenvolvimento de Software do Android, que é um conjunto de ferramentas que os desenvolvedores utilizam para criar aplicativos para *smartphones* e *tablets* com o sistema operacional Android. A IDE utilizada foi o Android Studio, uma vez que ela fornece uma interface gráfica que permite aos desenvolvedores executar tarefas de desenvolvimento mais rapidamente. O aplicativo foi implementado em incrementos, baseado nas metodologias ágeis. Em cada incremento foi construída uma nova funcionalidade, acompanhada da definição de testes.

Depois da construção de uma versão estável do aplicativo, foi realizada a fase de testes. Nessa fase, foi disponibilizada uma versão de testes em laboratório, utilizando possíveis usuários presentes no Instituto Federal Goiano - Campus Urutaí. Isso possibilitou a detecção de erros e de possíveis melhorias para garantir o perfeito funcionamento e aceitação do aplicativo. Por fim, o aplicativo foi disponibilizado para download na loja de aplicativos do Android.

4. Resultados e Discussões

4.1. Prototipação e validação das telas com Figma

Para prototipar e validar as telas foi utilizado o software Figma, uma ferramenta que nos possibilitou elaborar as telas a serem desenvolvidas no aplicativo. Estas telas foram criadas a partir dos requisitos citados na seção anterior, onde todas elas representam determinados fluxos. Neste aplicativo foi possível determinar os principais fluxos, sendo eles: cadastro ONG, cadastro dados bancários, criação e edição de caso, doação e confirmação de doação.

A Figura 2 mostra as telas presentes no fluxo que uma ONG terá que realizar para criar um caso para um animal.

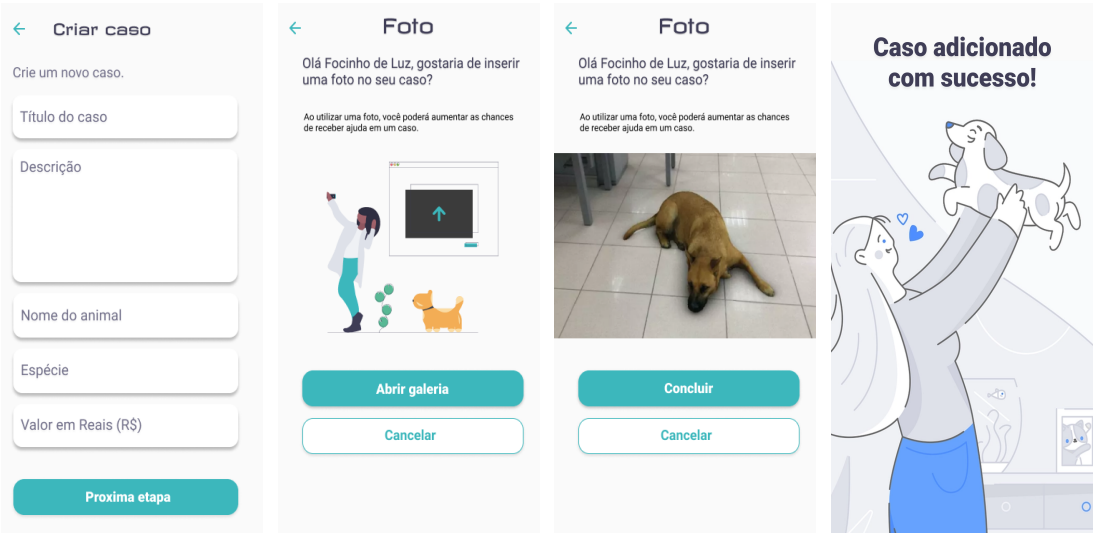


Figura 2. Fluxo para criação de um caso para um animal.

A primeira tela presente na Figura 2a é a responsável por coletar os dados do caso como titulo, descrição, nome, espécie e valor. Após esta etapa temos a Figura 2b e 2c que são as telas relacionadas ao envio de imagens do animal em questão. Por fim, temos a Figura 2d que confirmará a criação do caso.

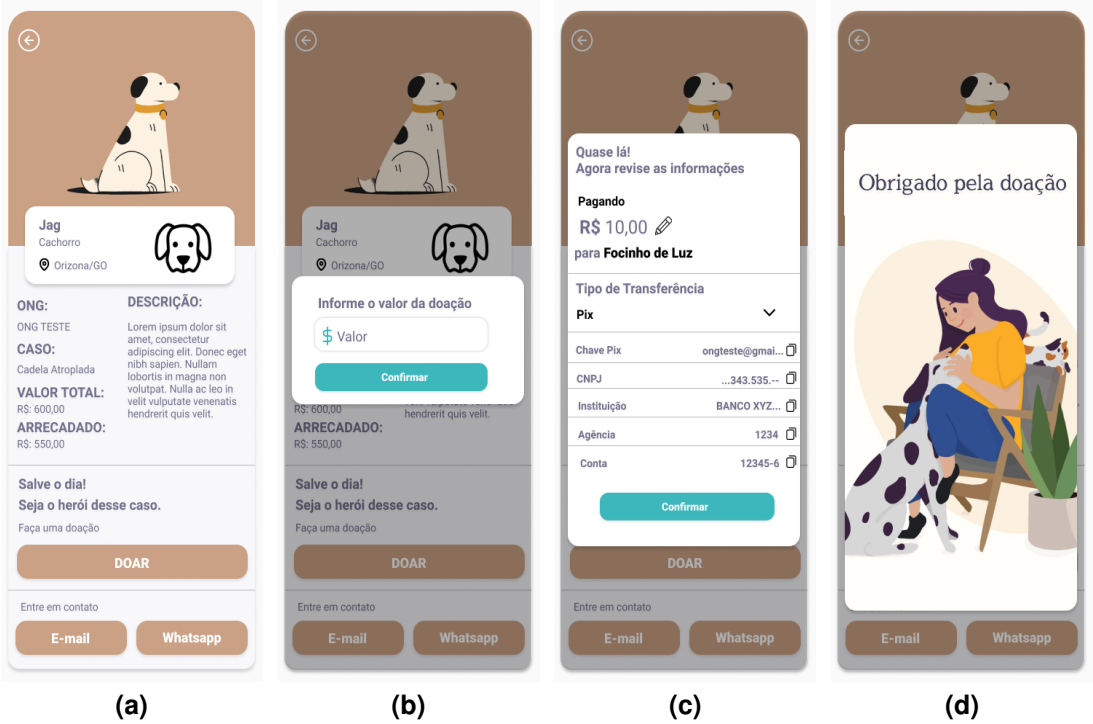


Figura 3. Fluxo para realizar doação em um caso de animal.

Outro exemplo de telas de um fluxo é exibido na Figura 3, onde são representados os passos necessários para realizar uma doação em um caso de animal. A Figura 3a apresenta a tela que mostra todos detalhes de um caso específico e ao acionar o botão doar a tela presente na Figura 3b será exibida para informar o valor a ser doado, e a tela da Figura 3c confirmará todas informações importantes para realizar a doação. Após todos estes passos será exibido a Figura 3d com a finalidade de agradecer o apoiador pela doação.

4.2. Modelagem da base de dados

Uma das etapas mais importantes durante o desenvolvimento de um aplicativo é a modelagem do banco de dados, pois nele será armazenado todas informações importantes para o seu funcionamento. A Figura 4 mostra a modelagem do banco de dados utilizando a notação Entidade Relacionamento (ER) pé de galinha que conta com um formato gráfico intuitivo.

A ONG contém o armazenamento de todas suas informações, com exceção da senha, que será gerenciada pelo *Google Authentication*. A ONG poderá possuir 0 ou vários casos de animais, e também possuirá 1 ou mais dados bancários para receberem as doações dos casos. Já o caso, além de suas informações, será armazenado as doações realizadas pelos apoiadores.

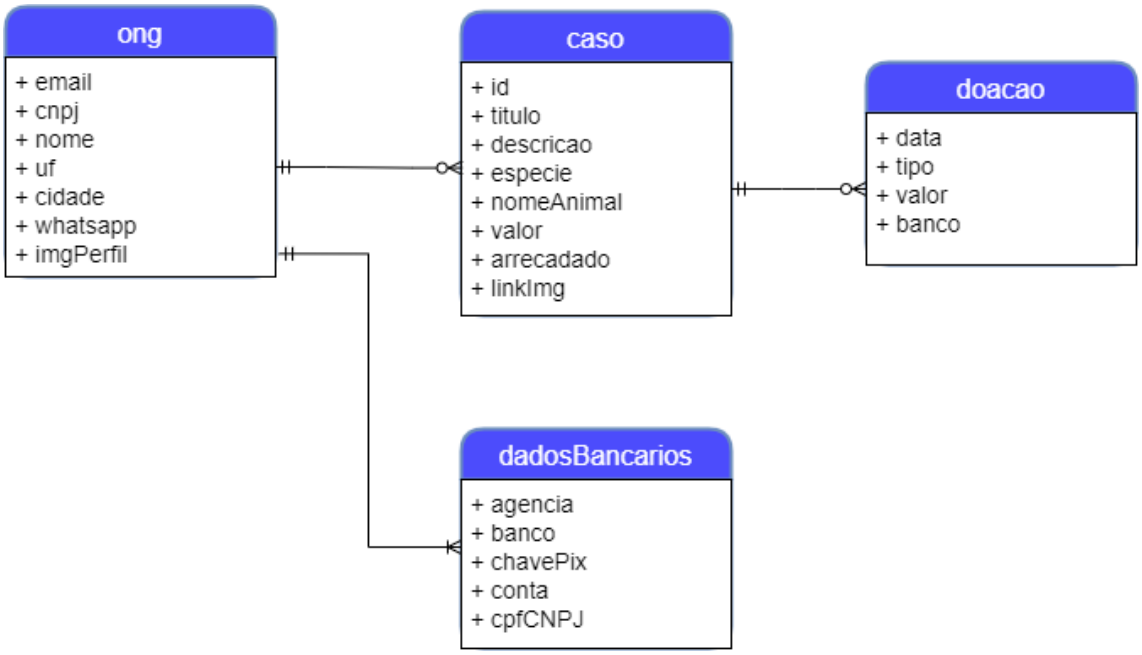


Figura 4. Modelagem do banco de dados.

No *Cloud Firestore* a organização foi realizada por meio de coleções, subcoleções e documentos. Um documento de ONG estará dentro de uma coleção de ONGs, e seus casos serão armazenados em subcoleções. Por fim, as doações estarão presentes como subcoleções nos documentos de casos.

4.3. Implementação

Após realizar a prototipação e a modelagem houve o inicio da etapa de implementação, onde inicialmente foi organizado a estrutura do projeto seguindo a arquitetura MVC. Posteriormente, desenvolveu-se todas as telas em XML se baseando nos protótipos dos fluxos de telas que foram explicados na Seção 4.1.

Com as telas construídas, foi realizado a conexão com o banco de dados *Cloud Firestore*, e as demais etapas focaram em vincular as telas com a parte lógica escrita na linguagem de programação Java. A Figura 5 exibe a organização dos pacotes do aplicativo.

O código deste aplicativo está fornecido gratuitamente em um repositório GitHub[1], que além de códigos, há também informações de todas tecnologias utilizadas e instruções de como executar este projeto. Por fim o aplicativo foi disponibilizado para *download* na loja de aplicativos Android (*Play Store*)[2].

[1] `https://github.com/souzavaltenis/Ipet-1`
[2] `https://play.google.com/store/apps/details?id=com.bdtgo.ipet`

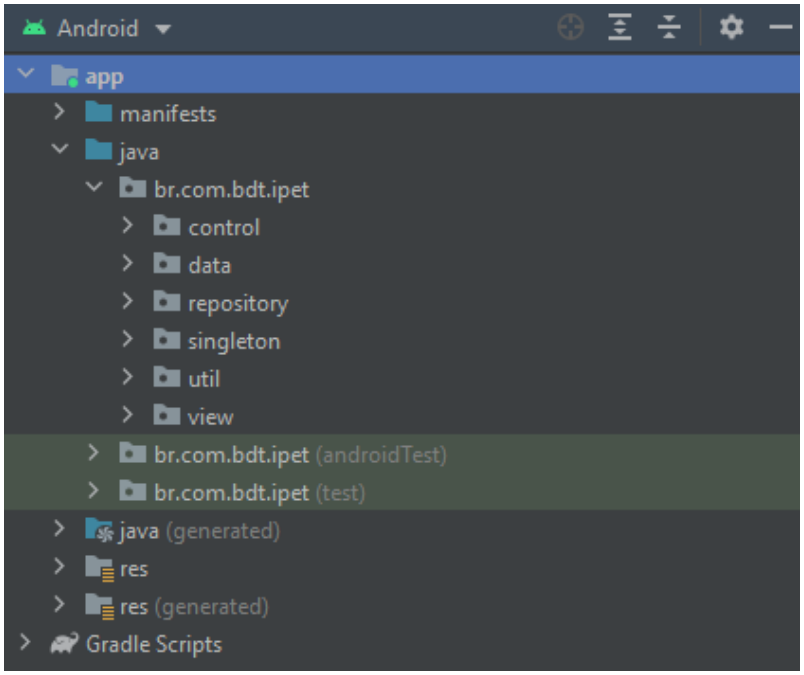


Figura 5. Organização dos pacotes de código do aplicativo.

5. Conclusão

Neste trabalho, mostramos a elaboração e o desenvolvimento de um aplicativo destinado a *smartphones* Android utilizando a linguagem de programação Java para auxiliar ONGs protetoras de animais. O aplicativo consegue trazer diversos benefícios paras as ONGs protetoras de animais, como por exemplo: criar casos para ajuda financeira ou adoção de um ou vários animais, atrair novos interessados em doar e controlar doações.

Já para os apoiadores, é possível encontrar casos de todo Brasil ou até os mais próximos por meio de filtragem, conseguir obter informações sobre a ONG, e possibilitar doações sem um valor mínimo em um determinado caso. Por fim, disponibilizamos de forma gratuita na loja de aplicativos Android (*Play Store*) a fim de alcançar o maior número de usuários e assim aumentar as conexões de ONGs protetoras de animais com apoiadores que possuam interesse em ajudar.

Como trabalho futuro poderá ser integrado um sistema de pagamentos *online* para facilitar as doações e a criação de notificações para novos casos de animais. Além disso, para ampliar o acesso de usuários, uma versão multiplataforma pretende ser desenvolvida e disponibilizada também para *smartphones* com sistema operacional IOS.

Referências

AdotaPet (2021). Adota pet go - adote um animal próximo a você. Disponível em: https://play.google.com/store/apps/details?id=com.labup.adotapetv2. Acessado em: 28 Abr 2021.

Deitel, H. M., Deitel, P. J., and Furmankiewicz, E. (2008). *Java: como programar*. Pearson educacion.

Firestore, C. (2021). Documentação firebase - firestore. Disponível em: https://firebase.google.com/docs/firestore. Acessado em: 06 Mar 2021.

Gomez, P., Casallas, R., and Roncancio, C. (2016). *Data schema does matter, even in NoSQL systems!* IEEE.

Instituto Pet Brasil (2019). Censo pet: 139,3 milhões de animais de estimação no brasil. Disponível em: http://institutopetbrasil.com/imprensa/censo-pet-1393-milhoes-de-animais-de-estimacao-no-brasil/. Acessado em: 05 Abr 2021.

Kumar, J. and Garg, V. (2017). *Security analysis of unstructured data in NOSQL MongoDB database*. IEEE.

Lemos, S. (2021). Cresce o número de adoções e de abandono de animais na pandemia. Disponível em: https://jornal.usp.br/atualidades/cresce-o-numero-de-adocoes-e-de-abandono-de-animais-na-pandemia/. Acessado em: 07 Abr 2021.

MEAUDOTE (2021). Um pouco sobre o meaudote. Disponível em: https://www.site.meaudote.com.br/sobre-o-meaudote. Acessado em: 27 Abr 2021.

Pereira, L. C. O. and da Silva, M. L. (2009). *Android para desenvolvedores*. Brasport.

PetPonto (2021). Petponto, app de adoção de animais. Disponível em: http://www.petponto.com/. Acessado em: 27 Abr 2021.

Sebrae (2017). Tudo sobre organizações não governamentais (ongs). Disponível em: https://www.sebrae.com.br/sites/PortalSebrae/artigos/artigoshome/o-que-e-uma-organizacao-nao-governamental-ong,ba5f4e64c093d510VgnVCM1000004c00210aRCRD. Acessado em: 10 Abr 2021.

Statcounter (2021). Operating system market share worldwide. Disponível em: https://gs.statcounter.com/os-market-share/mobile/brazil. Acessado em: 19 Out 2021.

Tubaldini, R. (2019). Ong de animais. Disponível em: https://www.cachorrogato.com.br/cachorros/ong-animais/. Acessado em: 16 Abr 2021.

TV TEM (2021). Ongs que cuidam de animais abandonados pedem ajuda em itapetininga. Disponível em: https://g1.globo.com/sp/sorocaba-jundiai/mundo-pet/noticia/2021/04/15/ongs-que-cuidam-de-animais-abandonados-pedem-ajuda.ghtml. Acessado em: 25 Abr 2021.

Velasco, C. (2019). Brasil tem mais de 170 mil animais abandonados sob cuidado de ongs, aponta instituto. Disponível em: https://g1.globo.com/sp/sao-paulo/noticia/2019/08/18/brasil-tem-mais-de-170-mil-animais-abandonados-sob-cuidado-de-ongs-aponta-instituto.ghtml. Acessado em: 22 Abr 2021.

Criação de um corpus para análises líricas de músicas brasileiras

Luiz Eduardo Gonçalves Silva[1], **Márcio de Souza Dias**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Catalão (UFCAT)

Abstract. *In Natural Language Processing (NLP), a corpus is a resource widely used in computational studies and relations of language phenomena. However, it is necessary that the corpus is well structured, organized and has enough content to train Machine Learning algorithms, for example, in order to obtain good results. This article presents the process of creation, construction and cleaning (removal of noise, normalization of sentences, among others) of a corpus aimed at lyrical analysis of Brazilian music, in addition to showing how its content was organized and presenting statistics about its content . For this, a free API provided by the Vagalume website, one of the most popular Brazilian song lyrics sites, is used to obtain the lyrics of the songs. As a result, a corpus composed of more than 150 thousand lyrics of different Brazilian musical genres is formed. Therefore, the corpus can be used by NLP and Machine Learning algorithms, which will learn lyrical patterns in Brazilian musical genres through the corpus, so that it is possible to classify them without the need for a rhythmic and sound evaluation.*

Resumo. *Em Processamento de Linguagem Natural (PLN), um corpus é um recurso muito utilizado em estudos e relações de fenômenos da língua de forma computacional. Entretanto, é necessário que o corpus esteja bem estruturado, organizado e que possua conteúdo suficiente para treinar algoritmos de Machine Learning, por exemplo, a fim de obter bons resultados. Este artigo apresenta o processo de criação, construção e limpeza (remoção de ruído, normalização das sentenças, dentre outros) de um corpus voltado para análises líricas de músicas brasileiras, além de mostrar de que forma seu conteúdo foi organizado e apresentar estatísticas sobre seu conteúdo. Para isso, uma API gratuita disponibilizada pelo website Vagalume, um dos sites brasileiros de letras de músicas mais populares, é utilizada para obter as letras das músicas. Como resultado, é formado um corpus composto de mais de 150 mil letras de músicas de diversos gêneros musicais brasileiros. Portanto, o corpus pode ser utilizado por algoritmos de PLN e Machine Learning, os quais vão aprender através do corpus padrões líricos nos gêneros musicais brasileiros, de forma que seja possível classificá-los sem a necessidade de uma avaliação rítmica e sonora.*

1. Introdução

A quantidade crescente de músicas disponíveis na internet exige ferramentas inteligentes para navegar e pesquisar em bancos de dados de música. Os sistemas de recomendação musicais presentes em aplicativos e websites voltados para o mercado de *streaming* de

música, podem ajudar o usuário a encontrar músicas de sua preferência. Esse processo tipicamente necessita de uma análise automática, i.e, classificação de acordo com o gênero, conteúdo, ou similaridade entre as músicas. Embora isso possa ser alcançado realizando uma análise sonora das músicas, conforme Juliano Henrique et al. (2020) apresentam no artigo *Texture selection for automatic music genre classification*, também é possível realizar tais validações utilizando somente a letra da música, conforme demonstrado por Fell e Sporleder (2014).

Entretanto, conforme afirmam Bertin-mahieux et al. (2011), a falta de dados pode ser um problema comum nos campos de análises estatísticas. Esse problema ainda é agravado quando se trata de músicas, devido às licenças musicais. Além disso, autores de algoritmos que utilizam de tais dados buscam fontes de dados confiáveis, de preferência bem organizados. Portanto, é notável a necessidade de criação ou utilização de um corpus em atividades que realizam processamento de grandes quantidades de informação. Isso é ainda mais importante em Processamento de Linguagem Natural, visto a necessidade de treinar algoritmos de Machine Learning para realização das análises dos fenômenos linguísticos, por exemplo.

Os artigos de Fell e Sporleder (2014) e Patrik Guimarães et al. (2020) demonstram a necessidade de criação de um corpus para realizar as análises propostas em seus respectivos artigos. O primeiro busca determinar os gêneros musicais, distinguir as melhores e as piores músicas e determinar o tempo aproximado de publicação de uma música utilizando as letras de música presentes no corpus criado por eles. O segundo tem um objetivo semelhante: detecção de gêneros musicais brasileiros utilizando as letras de músicas, também presentes no corpus criado pelos autores.

Desse modo, o objetivo desse trabalho é apresentar os passos necessários na construção de um corpus de letras de músicas brasileiras, detalhando organização e limpeza do corpus, de modo que algoritmos possam utilizá-lo para analisar letras de músicas brasileiras, evitando a necessidade de uma análise sonora. Na sessão de Fundamentação Teórica, o leitor será familiarizado com os termos utilizados no decorrer do artigo. A sessão de trabalhos relacionados apresenta artigos de outros autores que fizeram uma pesquisa semelhante ao trabalho realizado nesse artigo. A sessão de Metodologia irá apresentar a metodologia utilizada para alcançar o objetivos propostos nesse artigo. Por fim, a sessão de Conclusão apresenta um resumo do que foi alcançado no artigo, possíveis aplicações e trabalhos futuros.

2. Fundamentação Teórica

Nesta sessão, serão apresentados termos e definições utilizados no decorrer do artigo os quais facilitarão o entendimento da abordagem empregada neste trabalho.

Inicialmente, é necessário entender que um corpus é uma coleção de um grande número de textos armazenados em um dispositivo de computação [Wołk and Wołk 2017]. Este corpus será pré-processado (removendo ruídos, letras repetidas, normalização da terminação das sentenças), conforme explicado nas seções abaixo. O termo sentença é geralmente definido como uma palavra ou grupo de palavras que expressa uma ideia completa ao dar uma declaração/ordem, ou fazer uma pergunta, ou exclamar algo [LearnEnglish.net]. O corpus criado neste artigo será composto por letras de músicas recuperadas da internet, e será organizado em **gêneros musicais**. Gênero musical é um

conjunto de eventos musicais (reais ou possíveis) cujo curso é regido por um conjunto definido de regras socialmente aceitas, onde um evento musical pode ser definido como ”qualquer tipo de atividade realizada em torno de qualquer tipo de evento envolvendo som”[Fabbri 1982], ou seja, uma categoria convencional que identifica algumas peças musicais como pertencentes a uma tradição compartilhada ou conjunto de convenções.

Para obter as letras de música, o website Vagalume [1] será utilizado. O site Vagalume é um dos sites de música mais acessados em todo o Brasil e Portugal [Marktest 2009]. Ele fornece uma API [2] que será usada para recuperar as letras da internet.

Um *script* será usado para recuperar dados da internet. *Script* refere-se a um programa de computador que será executado e realizará uma ação, como: buscar músicas na internet e salvá-las no disco rígido, contabilizar a quantidade de artistas no corpus, entre outras. Para criar os *scripts*, duas linguagens de programação principais serão usadas: Javascript e Python. JavaScript (frequentemente abreviado como JS) é uma linguagem de programação leve, interpretada e orientada a objetos com funções de primeira classe, conhecida como a linguagem de *scripting* para páginas Web, mas também utilizada em muitos ambientes fora dos navegadores. Ela é uma linguagem de *scripting* baseada em protótipos, multi-paradigma e dinâmica, suportando os estilos orientado a objetos, imperativo e funcional [Mozilla 2021]. Já Python é uma linguagem de programação interpretada, orientada a objetos e de alto nível com semântica dinâmica. Suas estruturas de dados embutidas de alto nível, combinadas com tipagem dinâmica e vinculação dinâmica, tornam-no muito atraente para o desenvolvimento rápido de aplicativos, bem como para uso como linguagem de *script* ou cola para conectar componentes existentes [Python Software Foundation 2021].

As letras de música serão salvas no formato JSON (*JavaScript Object Notation* - Notação de Objetos JavaScript), que é uma formatação leve de troca de dados. Para seres humanos, é fácil de ler e escrever. Para máquinas, é fácil de interpretar e gerar. Está baseado em um subconjunto da linguagem de programação JavaScript, Standard ECMA-262 3a Edição -Dezembro - 1999. JSON é em formato texto e completamente independente de linguagem, pois usa convenções que são familiares às linguagens C e familiares, incluindo C++, C#, Java, JavaScript, Perl, Python e muitas outras. Estas propriedades fazem com que JSON seja um formato ideal de troca de dados. [json.org 2017].

Finalmente, o termo *webscrapping* refere-se a uma técnica a qual um programa de computador extrai dados de um website da internet para uma saída legível [Martin Perez 2021].

3. Trabalhos Relacionados

O artigo de Michael Fell e Caroline Sporleder (2014) apresenta uma nova abordagem para analisar e classificar letras de música, experimentando tanto modelos de n-gram quanto recursos mais sofisticados que modelam diferentes dimensões de um texto de música, como vocabulário, estilo, semântica, orientação para o mundo e a estrutura da música. É mostrado que eles podem ser combinados com recursos de n-gram para obter ganhos de

[1] https://www.vagalume.com.br/

[2] https://api.vagalume.com.br/

desempenho em três tarefas de classificação diferentes: detecção de gênero, distinguir as melhores e as piores músicas e determinar o tempo aproximado de publicação de uma música. Para alcançar o objetivo do artigo, os autores criaram um corpus com mais de 400 mil músicas em inglês, contendo letras de músicas de mais de 7 mil e 200 artistas [Fell and Sporleder 2014].

O artigo de Paula Cardoso et al. (2011) apresenta o CSTNews, um corpus com anotações de discurso para fomento à pesquisa em sumarização de documentos únicos e múltiplos. O corpus compreende 50 grupos de textos de notícias em português do Brasil e alguns materiais relacionados, que inclui um conjunto de resumos manuais de um único documento e um conjunto de resumos manuais e automáticos de vários documentos. O corpus CSTNews é composto por 50 grupos de textos de notícias coletados em 2007. Eles abordam diversos assuntos de agências de notícias online populares no Brasil, a saber, Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. Cada cluster veicula de 2 a 3 textos redigidos em português brasileiro, coletados de acordo com sua repercussão na época em que eles foram publicados. O corpus resume ao todo 140 textos, totalizando 2.088 frases e 47.240 palavras. Em média, o corpus veicula 2,8 textos, 41,76 sentenças e 944,8 palavras por cluster [Cardoso et al. 2011].

O trabalho de Knees et al. (2005) apresenta uma abordagem para a extração de letras de música da internet. Como as letras encontradas na internet apresentam erros, como erros de digitação, são utilizadas várias versões da mesma letra obtidas de diferentes fontes (sites). Isso é obtido por meio do Alinhamento de Sequências Múltiplas. Esta técnica é emprestada da Bioinformática, onde é usada para alinhar sequências de DNA e proteínas. Para atingir os objetivos propostos no artigo, Knees et al. o utiliza para encontrar sequências de palavras quase ideais correspondentes em todas as páginas de letras [Knees et al. 2005]. Sites distintos são alinhados e examinados para encaixar sequências de palavras, encontrando as partes que têm o maior consenso entre as diferentes fontes.

4. Metodologia

A metodologia utilizada para a realização do trabalho é sustentada nos seguintes pilares: coleta de músicas para a composição do corpus e limpeza do corpus para eliminação de ruído (letras repetidas, artistas repetidos, normalização da terminação das letras, agrupamento de letras por gênero, etc).

4.1. Coleta de dados

A coleta de dados foi o primeiro passo realizado para iniciar a construção do corpus. Diversos websites de letras de música estão disponíveis na internet. Entretanto, não era desejável utilizar um webiste onde fosse necessário realizar *webscrapping* para obter as letras, visto que o código fonte do site pode mudar, exigindo que o *script* que realiza a coleta de dados seja alterado de modo a permanecer utilizável. Portanto, era desejável encontrar um website que fornecesse uma API para obtenção das letras de música, visto que uma API já fornece uma interface prática e mais legível para uso humano.

Após uma busca na internet, a API disponibilizada pelo webiste Vagalume foi encontrada. Visto que o site possui um grande acervo de letras de músicas brasileiras disponíveis, foi definido que o mesmo seria utilizado. Portanto, a documentação da API foi estudada a fim de que a mesma pudesse ser utilizada. Outras APIs foram encontradas na

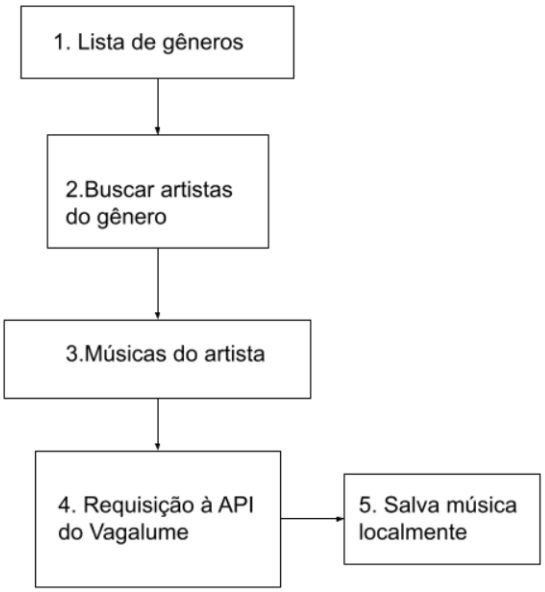


Figura 1. Fluxo de coleta de dados

internet, porém ou eram pagas ou não possuíam um acervo de letras de músicas brasileiras tão grande quanto a API do Vagalume.

O segundo passo para a coleta de dados foi escrever um *script* que realizasse as chamadas à API do Vagalume, recuperando a letra da música e salvando-a localmente. O *script* seguia o seguinte fluxo, conforme a figura 1:

1. Uma lista de gêneros foi escolhida. Essa lista foi obtida analisando listas na internet de gêneros mais populares no Brasil, como mostra Rodrigo Ortega (2020), por exemplo. A lista final possuía os seguintes 15 gêneros: axé, bossa nova, forró, funk, funk carioca, gospel, mpb, pagode, pop, pop rock, rap, regional, rock, samba, sertanejo.
2. Para cada gênero, um *scrapping* foi realizado para pegar todos os artistas através do website do Vagalume [3], visto que a API não disponibilizava os artistas por gênero. Os artistas eram obtidos e armazenados, também, em uma lista durante a execução do *script*.
3. Em seguida, para cada artista de cada gênero, uma segunda página do site Vagalume foi utilizada [4], a qual permitia recuperar o nome de todas as músicas de cada artista, as quais foram armazenadas em uma lista durante a execução do *script*.
4. Por fim, para cada nome de música do artista, uma requisição foi realizada à API do vagalume [5] para obter a letra da música. A letra era, então, armazenada no diretório correto, seguindo o padrão: Genders/GÊNERO/ARTISTA/NOME DA MUSICA.json. [6], onde a pasta Genders é a pasta raiz que contém todo o corpus.

Nota-se que entre cada requisição, um intervalo de dez segundos foi realizado, de

[3]Exemplo da URL utilizada para o gênero axé: https://www.vagalume.com.br/browse/style/axe.html.

[4]Exemplo da segunda página acessada: https://www.vagalume.com.br/netinho/index.js

[5]https://api.vagalume.com.br/

[6]Um exemplo para o artista Alexandre Peixe, do gênero Axé, música Abainar: Genders/axe/Alexandre Peixe/Abainar.json

modo que a API do Vagalume não recebesse um número muito grande de requisições e bloqueasse as chamadas. A execução do script durou mais de 48 horas, com intervalos intermitentes.

4.2. Limpeza do corpus

Após a coleta de dados e organização do mesmo, o segundo passo para a criação do corpus foi a limpeza dos dados. Conforme Knees et al (2005), letras de músicas extraídas da internet necessitam ser sanitizadas, i.e, é necessário excluir letras repetidas, ou remover estruturas que não pertencem à letra original, como [REFRÃO], x2, x3, por exemplo, uma vez que as mesmas são publicadas por diferentes usuários da internet, os quais podem interpretar de maneira diferente dada palavra ou sentença da música, por exemplo.

Além disso, estruturas da letra podem ser representadas de diferentes maneiras. Um exemplo disso é a repetição de uma dada estrutura da letra, como o refrão. Alguns autores escolhem a utilização de metadados para representar dada estrutura quando a mesma se repete. Uma possível maneira de escrever seria a utilização de abreviações como (x2), (x3) ou a palavra REFRÃO, funcionando como um símbolo que deve ser substituído pela letra do refrão.

Desse modo, alguns *scripts* de normalização e sanitização foram escritos. O fluxo para limpeza segue como abaixo:

- A princípio, verificou-se que alguns gêneros, normalmente semelhantes, possuíam o mesmo artista. Isso é esperado uma vez que alguns artistas podem experimentar diferentes gêneros musicais. Entretanto, não seria viável escrever um *script* que realizasse a validação do gênero do artista automaticamente, e considerando o baixo número de casos, optou-se por um trabalho manual, o qual consistia em procurar pelo artista na internet, verificar em diferentes fontes qual a classificação dada para seu estilo de música e, finalmente, manter o diretório do gênero predominante, removendo os duplicados. Posteriormente, letras de música repetidas foram excluídas.
- A seguir, para cada letra de música de cada artista do corpus, inicialmente fora removidos metadados de repetição, como: (x2), x3, REFRÃO, dentre outros.
- Em seguida, foi realizada a normalização das terminações das sentenças. Assim, todas as sentenças de cada uma das letras de música do corpus são finalizadas com um ponto final. A realização desse passo permite que usuários que utilizem o corpus executem algoritmos de *chunk* e *pos tag* de maneira facilitada, visto que o fim da sentença é normalizado.
- Dado que o objetivo do artigo é construir um corpus composto exclusivamente de letras de música brasileiras, foi realizada uma análise utilizando bibliotecas de linguagem [7] para validar o idioma das letras. Esse passo foi facilitado visto que a API do Vagalume já classificava o idioma das letras. Músicas que não fossem do português brasileiro foram excluídas.

4.3. Estatísticas do corpus

Após a coleta e limpeza do corpus, dados estatísticos foram obtidos, com a finalidade de analisar algumas dados, como quantidade de artistas, média de sentenças, dentre outras

[7] https://pypi.org/project/langdetect/

métricas, a partir das quais é possível constatar qual gênero possui maior quantidade de sentenças, por exemplo, e o que isso implica na análise realizada no trabalho.

O corpus criado consiste de um total de 1310 artistas nos quinze gêneros a seguir (número de artistas por gênero): Axé: 40, Bossa Nova: 25, Forró: 102, Funk/Funk Carioca: 68,Gospel: 366, MPB: 114, Pagode: 52, Pop: 56, Pop Rock: 36, Rap: 88, Regional: 15,Rock: 76, Samba: 83, Sertanejo: 189. Na tabela 1 é apresentado alguns dados estatísticos obtidos referentes ao corpus.

Tabela 1. Dados estatísticos do corpus

	Total
Total de gêneros	15
Total de artistas	1310
Média de sentenças por gênero	529.939
Média de palavras por gênero	1.254.378
Média de letras por gênero	5.054.670
Total de sentenças no corpus	7.949.099
Total de palavras no corpus	18.815.683
Total de letras no corpus	75.820.061

A partir dos dados coletados, foi possível perceber que: para o corpus criado, o gênero Gospel é o que possui a maior quantidade de sentenças, ganhando por pouco do gênero Sertanejo e do Rap. Essa informação implica que tais gêneros possuem uma maior quantidade de letra por música e uma maior quantidade de artistas por gênero. Foi obtido também a informações de que o gênero Funk possui a menor quantidade de sentenças dentre todos os gêneros.

5. Conclusão

Portanto, o artigo apresentou a criação de um corpus de letras de músicas brasileiras, desde a coleta de dados até a limpeza do mesmo. Logo, pode-se considerar que os objetivos propostos foram alcançados, i.e, disponibilizar um corpus com grande quantidade de dados, bem organizado e normalizado. Possíveis aplicações para o corpus envolvem a utilização de algoritmos de Machine Learning para análise lírica das músicas. O corpus também pode ser utilizado por algoritmos de PLN e Machine Learning, como um conjunto de dados de treinamento de modelos, os quais vão aprender através do corpus padrões líricos nos gêneros musicais brasileiros, de forma que seja possível classificá-los sem a necessidade de uma avaliação rítmica e sonora.

Referências

Bertin-mahieux, T., Ellis, D. P. W., Whitman, B., and Lamere, P. (2011). The million song dataset. In *In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR*.

Cardoso, P. C. F., Maziero, E. G., Jorge, M. L. R. C., Seno, E. M. R., Felippo, A. D., Rino, L. H. M., Nunes, M. d. G. V., and Pardo, T. A. S. (2011). Cstnews: a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Brazilian Symposium in Information and Human Language Technology - STIL*. SBC.

Fabbri, F. (1982). A theory of musical genres: Two applications.

Fell, M. and Sporleder, C. (2014). Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 620–631, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Foleiss, J. H. and Tavares, T. F. (2020). Texture selection for automatic music genre classification. *Applied Soft Computing*, page 106127.

json.org (2017). Introducing json. `https://www.json.org/json-pt.html`. [Online; accessed 19-October-2021].

Knees, P., Schedl, M., and Widmer, G. (2005). Multiple lyrics alignment: Automatic retrieval of song lyrics. pages 564–569.

LearnEnglish.net. Sentence: Definition and types: Learn english.

Marktest, G. (2009). Vagalume lidera sites de música.

Martin Perez (2021). What is web scraping and what is it used for? `https://www.parsehub.com/blog/what-is-web-scraping/`. [Online; accessed 19-October-2021].

Mozilla (2021). Sobre javascript. `https://developer.mozilla.org/pt-BR/docs/Web/JavaScript/About_JavaScript`. [Online; accessed 19-October-2021].

Python Software Foundation (2021). What is python? executive summary. `https://www.python.org/doc/essays/blurb/`. [Online; accessed 19-October-2021].

Rodrigo Ortega (2020). Forró cresce no streaming e supera audição de rap e pop nacional, puxado pela pisadinha. `https://g1.globo.com/pop-arte/musica/noticia/2020/12/26/forro-cresce-no-streaming-e-supera-audicao-de-rap-e-pop-nacional-puxa` `ghtml?utm_source=twitter&utm_medium=social&utm_campaign=g1`. [Online; accessed 19-October-2021].

Rogério Theodoro de Brito (2001). Alinhamentos de múltiplas sequências. `https://www.ime.usp.br/~rbrito/docs/quali-texto.pdf`. [Online; accessed 19-October-2021].

Wołk, K. and Wołk, A. (2017). Automatic parallel data mining after bilingual document alignment. pages 317–327.

Simulação do processo de evacuação de pedestres no restaurante universitário da Universidade Federal de Catalão via Autômatos Celulares

Matheus Matos Machado [1], **Sérgio Francisco da Silva** [1]

[1]Instituto de Biotecnologia – Departamento de Ciência da Computação
Universidade Federal Catalão (UFCat)
CEP: 75704-020 – Av. Dr. Lamartine Pinto de Avelar, 1120 – Catalão – GO – Brazil

m.matos1012.m@gmail.com

sergio@ufcat.edu.br

Abstract. *Pedestrian evacuation time analysis for a given building environment is a vital part of their planning process. This paper aims to simulate the evacuation of pedestrians in an environment similar to the university restaurant at the Federal University of Catalão. The simulations are carried out using a cellular automaton, 100 simulations are carried out for each of the five different environment capacities, totaling 500 simulations. It analyzes the total evacuation time and the number of collisions that occur during the evacuation process. At the end of this project, it was possible to analyze an increase in the average evacuation time, but lower than the proportion in which the number of diverse people, while the number of collisions was much above this proportion.*

Resumo. *A análise do tempo de evacuação de pedestres, para um dado ambiente de uma construção, é parte vital do processo de planejamento dele. Este artigo visa simular a evacuação de pedestres em um ambiente similar ao restaurante universitário da Universidade Federal de Catalão. As simulações são realizadas utilizando um autômato celular, são realizadas 100 simulações para cada uma das cinco diferentes taxas de lotação do ambiente, totalizando assim 500 simulações. Visando analisar o tempo total de evacuação e a quantidade de colisões que ocorrem durante o processo de evacuação. Ao final desse projeto, foi possível analisar um aumento no tempo médio de evacuação, porém inferior à proporção em que o número de pessoas variou, enquanto o número de colisões ficou muito acima dessa proporção.*

1. Introdução

A população mundial vem aumentando ao longo dos anos, passando de uma estimativa de 2.526 bilhões de pessoas em 1950, para 4.449 bilhões no início da década de 1980 e chegando aos incríveis 7.713 bilhões de habitantes em 2019 [Nations 2019]. Segundo projeção publicada pela Organização das Nações Unidas dadas em [Nations 2019], espera-se que em 2050, 9,7 bilhões de pessoas estejam vivendo ao redor de todo o planeta. Este rápido crescimento da população mundial tem provocado grandes impactos em diversos aspectos da sociedade, até mesmo em questões primordiais para a vida, como saúde, educação, alimentação, acesso a água potável, entre outros.

Ao longo das últimas décadas tornou-se comum a migração de pessoas para grandes centros urbanos em busca de empregos, subsídios, acesso à educação, entre outros [Alves et al. 2011]. Porém, com este ajuntamento de pessoas nos centros urbanos, diversos desafios têm sido enfrentados no planejamento de infraestruturas de moradias, transporte, saneamento e segurança pública. Com o perceptível inchaço demográfico que se observa na atualidade, é rotineira a aglomeração de pessoas em um mesmo local. Um fato que torna-se tão ou mais relevante do que viabilizar a chegada das pessoas a estes locais é a circunstância em que se faz necessário a evacuação do ambiente em detrimento de alguma emergência.

A estrutura de uma grande praça de eventos ou qualquer local que pretende receber um grande número de pessoas deve ser muito estudada pois, as disposições de obstáculos físicos, do imobiliários e das saídas podem ser determinantes no sucesso da evacuação de um local durante uma urgência [Feliciani et al. 2020]. Além do fator estrutural, o comportamento de uma multidão durante uma situação de emergência é um fato extremamente expressivo, visto que em situações desesperadoras não é atípico ouvirmos relatos de lesões ocasionadas por colisões e empurrões, e até mesmo esmagamentos provenientes de situações nas quais múltiplos pedestres dirigem-se para uma mesma porta de saída [Helbing and Johansson 2009].

Tendo em vista a importância dos pontos apresentados anteriormente, é natural e relevante, realizar simulações de evacuação de emergência em locais que sabidamente receberão um grande número de pessoas, antes da liberação do mesmo para entrar em funcionamento. Entretanto, efetuar experimentos práticos com humanos com este intuito não é algo simples de se fazer, pois os custos são altos, os dados experimentais são difíceis de se captar, e podem não corresponder à realidade pois fenômenos como pânico, quedas e choques são muito difíceis de serem simulados na prática [Alizadeh 2011]. Estes aspectos tornam atrativo a alternativa de se concretizar tais simulações utilizando métodos computacionais.

Este trabalho tem como objetivo apresentar um autômato celular, desenvolvido para simular o processo de evacuação do pedestres no restaurante universitário da Universidade Federal de Catalão. A partir deste autômato pode-se efetuar diversas simulações com diferentes capacidades, desde 25% até 100% de sua capacidade máxima onde, pode-se observar como o tempo de evacuação e quantidade de colisões aumenta de acordo com o aumento na lotação do ambiente .

O texto organiza-se da seguinte maneira: Primeiramente há a sessão de introdução, onde discorre-se sobre a necessidade das simulações de evacuação; em seguida apresenta-se a sessão de Visão geral de Dinâmica de evacuação, onde é discutido os principais aspectos da dinâmica de evacuação; a terceiro sessão consiste em uma fundamentação do teórica sobre os Autômatos celulares; a quarte sessão tem como foco descrever o modelo de autômato celular desenvolvido neste trabalho; a quinta sessão apresenta a metodologia utilizada para realizar os testes; a sexta sessão apresenta os resultados obtidos a partir dos testes; por fim, a sétima sessão é a conclusão do trabalho.

2. Visão geral de Dinâmica de Evacuação

Pedestres são objetos tridimensionais e uma completa descrição de seus movimentos é bastante difícil. Assim, normalmente, a dinâmica de evacuação de pedestres é tratada em

duas dimensões, considerando a projeção vertical do corpo [Schadschneider et al. 2011]. A seguir são descritos os principais fenômenos coletivos que ocorrem na evacuação de pedestres e que devem ser tratados por modelos de simulação efetivos. Posteriormente, são introduzidas as principais variáveis da dinâmica de evacuação de pedestres.

2.1. Fenômenos coletivos

Uma das razões da investigação de dinâmica de evacuação de pedestres ser desafiadora é a alta variedade de fenômenos coletivos e de auto-organização que podem ocorrer. Esses efeitos coletivos, chamados de macroscópicos, refletem as interações individuais, chamadas de microscópicas, e assim se tornam também uma importante fonte de informação de qualquer modelagem [Schadschneider et al. 2011, Helbing and Johansson 2009].

2.1.1. Gargalo

De acordo [Tanenbaum 2003], quando a carga oferecida a qualquer rede é maior que sua capacidade, acontece um congestionamento. A partir disso pode-se esboçar um paralelo com o processo de evacuação, pois quando uma passagem de um ambiente para outro não oferece a vazão apropriada para atender a quantidade de pedestres que estão chegando, ocorre um gargalo.

2.1.2. *Jamming*

Este fenômeno ocorre devido a alta densidade de pedestres em locais com fluxo insuficiente [Schadschneider et al. 2011]. Quando o fluxo é insuficiente, pode ocorrer um *jamming*, que consiste, basicamente, em um entrave no processo de evacuação, impedindo que mais pedestres passem pelo gargalo.

Este fenômeno não depende fortemente da dinâmica microscópica de cada pedestre. Ele uma uma consequência de um princípio de exclusão: o espaço ocupado por um pedestre não pode ser ocupado por outros. O fenômeno de *jamming* também ocorrer em situações de contra-fluxo, por exemplo quando grupo de pedestres movimenta em direção contrária a um outro grupo.

2.1.3. Situações de emergência e Pânico

Conforme [Schadschneider et al. 2011] em situações de emergência vários fenômenos coletivos têm sido reportados e, várias vezes, estes têm sido atribuídos erroneamente ao comportamento de pânico. Embora não há uma definição precisa do que é pânico, usualmente certos aspectos são associados a este conceito. Tipicamente pânico ocorre em situações onde pessoas competem por recursos escassos (como espaço seguro ou acesso a uma saída) que levam ao egoísmo, ou comportamento associal, ou até mesmo completamente irracional e o contagiante que afetam grandes grupos.

2.2. Variáveis fundamentais da Dinâmica de Evacuação

As variáveis normalmente observáveis em evacuação de pedestres são:

2.2.1. Fluxo

O fluxo está relacionado à capacidade do gargalo, a densidade de partículas (pedestres) e a velocidade das partículas [Schadschneider et al. 2011]. Na literatura há várias modelagens matemáticas de fluxo, envolvendo variados conceitos como hidrodinâmica, diagrama fundamental de fluxo de tráfego, entre outros.

2.2.2. Vazão do gargalo

A vazão do gargalo é dado pelo número de pedestres cruzando o gargalo por unidade de tempo. Conforme [Schadschneider et al. 2011], uma das mais importantes questões práticas de evacuação é como a capacidade de evacuação aumenta com o aumento a largura do gargalo. Ainda conforme apontado por [Schadschneider et al. 2011], a intuição de que o aumento da largura do gargalo aumenta proporcionalmente a capacidade de evacuação não é sempre verdadeira pois a evacuação depende de outras variáveis como a dispersão da multidão e o posicionamento do gargalo. Conforme reportado em [Hoogendoorn and Daamen 2005], situações com formação de linhas em um único sentido aumentam a capacidade de evacuação por resultar no que é chamado pelo autores de "efeito zíper".

2.2.3. Bloqueios em situações de competição

Por definição um gargalo é um recurso limitado e é possível que em situações de pessoas competindo pela saída, o fluxo seja diferente daquele em condições normais, podendo ocorrer até mesmo bloqueios. Bloqueio também pode ocorrer a nível das partículas em decorrência de colisões e pânico, onde a pessoa não sabe o que fazer e fica parada por um instante [Schadschneider et al. 2011].

2.2.4. Topografia

A topografia do terreno pode afetar a evacuação de diversas maneiras, ela pode ter sua eficiência prejudicada (o tempo total de evacuação pode ser maior do que a realizada em um terreno com condições normais) por certos fatores adversos, como por exemplo, a encosta de uma montanha, adrenalina, obstruções e se uma pessoa está subindo ou descendo. Devido a este tipo de influência topográfica do solo, é costume marcar as rotas por onde as pessoas devem caminhar, normalmente, em uma tentativa de diminuir a inclinação, formam um caminho em zigue-zague [Velasquez and Alvarez-Alvarado 2021].

2.2.5. Obstáculos

Conforme [Varas et al. 2007], a modelagem e estudo de evacuação com obstáculos é de fundamental importância, pois estes estão presentes em muitas situações reais de evacuações de emergência como salas, cinemas, aeronaves, restaurantes, etc. Além disso, em situações de emergência, os pedestres podem cair e se tornar obstáculos para outros

pedestres. Um problema relevante a este respeito é como distribuir obstáculos no ambiente para melhorar os tempos de evacuação [Feliciani et al. 2020].

3. Autômatos celulares

As origens dos autômatos celulares são muitas vezes associadas ao matemático húngaro John von Neumann, o qual tentava criar uma máquina auto-replicante auto-denominada autômato. A questão que ele aborda é a seguinte: "Pode-se construir um agregado de tais elementos de tal maneira que se colocado em um reservatório, onde todos estes elementos flutuam, no final cada um deles pode ser outro autômato exatamente igual ao original?" A partir disso Neumann começa a modelar um argumento para mostrar que isso é, a princípio, viável [Schiff 2007]. Sob influência de Stanislaw Ulan, ele arquitetou uma rede (*grid*) bidimensional de máquinas de estados finitas (*Finite-states machines* – FSM) interconectadas localmente, as quais foram nomeadas como células. Tais células poderiam oscilar entre 29 estados possíveis de forma síncrona. Um mesmo conjunto de regras locais rege todas as máquinas de estados presentes no *grid*, o que equipara este agrupamento homogêneo a sistemas físicos e biológicos [de Souza Rosa 2018].

3.1. Definição

Conforme [Wolfram 1982], um autômato celular consiste em uma sequência de locais, os quais contém os valores 0 ou 1, dispostos em uma linha, ou eixo. De acordo com um conjunto de regras definidas envolvendo os valores de seus "vizinhos" mais próximos, o valor em cada um dos locais evolui de modo determinístico com o tempo. Usualmente, os locais de um CA podem ser dispostos em qualquer rede (*lattice*) regular, e cada um dos locais pode assumir qualquer conjunto discreto de valores.

Qualquer sistema com muitos elementos discretos idênticos passando por interações locais determinísticas pode ser modelado como um autômato celular [Wolfram 1982]. Também é possível encontrar alguns exemplos de autômatos celulares não-triviais, obtidos quando a evolução local não é linear, como o crescimento de um floco de neve. Na teoria dos números também pode-se observar sistemas matemáticos que assemelham-se a autômatos celulares [de Souza Rosa 2018].

3.2. Geometria de Célula e de Látice

A geometria de um autômato celular é caracterizada por um conjunto de propriedades que inclui a dimensão da célula, seu formato e a organização destas em uma malha ou grade [de Souza Rosa 2018]. As dimensões de um autômato celular podem ser, unidimensional, onde este é representado por um vetor, bidimensional onde é representado por uma matriz, ou tridimensional. Grande parte dos trabalhos de evacuação de pessoas desenvolvidos utilizam modelos bidimensionais, porém, em cenários mais complexos podem ser utilizados modelos tridimensionais.

Quanto ao formato, as células do autômato celular podem assumir diversas formas geométricas, estas podendo ser triangulares, retangulares, hexagonais, entre outras. Deve-se ressaltar que, em um sistema de autômatos celulares, todas as células devem ter a mesma forma [de Oliveira Carneiro 2012].

3.3. Vizinhança

A função da vizinhança é estabelecer o conjunto de células vizinhas de cada célula que serão considerados na atualização de seu estado nos instantes de tempo subsequentes [de Oliveira Carneiro 2012]. A definição do tipo de vizinhança a ser escolhida para o autômato depende do tipo de problema a ser solucionado, levando em consideração suas características, escolhendo assim qual será a representação mais adequada para cada caso [Lima 2007]. Dentre os principais modelos de vizinhança destaca-se a de Von Neumann que, como pode-se observar na Figura 1, cada célula inter-relaciona-se com as células verticais e horizontais adjacentes, sendo uma célula em cada sentido, no modelo convencional, e duas células no modelo estendido, tal qual o modelo anterior. Outra vizinhança que se sobressai é a de Moore, a qual pode ser observado na Figura 2; nela cada célula relaciona-se tanto com as células horizontais e verticais adjacentes, quanto com as células horizontais, em seu modelo estendido seu raio de interação passa de uma para duas células.

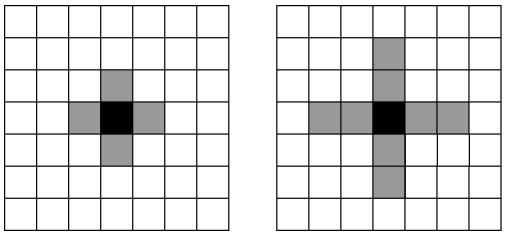


Figura 1. Vizinhança de Von Neumann e Von Neumann Estendida.

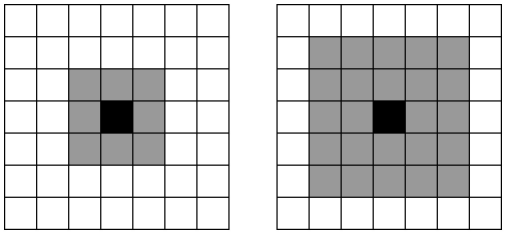


Figura 2. Vizinhança de Moore e Moore Estendida

3.4. Regras Locais e condições iniciais

Após definir a geometria, dimensão e vizinhança do autômato celular, deve-se definir qual será o estado inicial das células do autômato em questão, o qual pode ser aleatório ou construído singularmente para o problema a ser abordado. Após isso define-se quais serão as regras de transição, as quais são responsáveis por estabelecer o novo estado das células do autômato. A cada instante de tempo o autômato evoluirá mediante a aplicação das regras locais, o que resultará no comportamento idealizado para o autômato [de Souza Rosa 2018].

4. Descrição do modelo desenvolvido

O modelo de autômato celular bidimensional utilizado neste trabalho é baseado no modelo desenvolvido por [Burstedde et al. 2001]. As células presentes nesta implementação

possuem dimensões de 40×40 cm, a taxa de atualização do autômato é de 0,3s, três passos por segundo, resultando em uma velocidade aproximada de 1,3 m/s. A vizinhança utilizada é a de Moore, onde todas as células adjacentes a que se o pedestre encontra-se, são elegíveis para serem selecionadas durante a movimentação. O fator de pânico não é considerado nesta implementação.

4.1. Regras de Transição

A cada instante, cada partícula encolhe uma direção com base em uma matriz de preferência 3×3. Nesta matriz, a célula central representa a probabilidade do pedestre não se mover naquele instante; as oito células restantes representam a probabilidade de transição em cada direção. A cada instante, cada partícula tem acesso a essa matriz com base na célula que a partícula ocupa e no seus vizinhos. A cada instante de tempo, cada partícula escolhe a transição desejada de acordo com essas probabilidades. Se a célula alvo não estiver ocupada, a partícula se move para aquela célula. Se a célula alvo da partícula estiver ocupada, a partícula fica parada e conta-se uma colisão.

4.2. Campo de Piso

[Burstedde et al. 2001] introduziu o conceito de um campo de piso que é modificado pelos pedestres, que, por sua vez, modifica as probabilidades de transição. Este modelo permite levar em conta as interações entre os pedestres, a geometria do ambiente e obstáculos de um modo simples e unificado, sem perder as vantagens das regras de transição locais.

De modo geral, um campo de piso pode ser estático ou dinâmico. Um campo de piso estático S não evolui com o tempo e não muda na presença de pedestres. Este tipo de campo de piso pode ser usado para especificar regiões do ambiente que são mais atrativas, como saídas de emergência ou portas. Em contraste um campo de piso dinâmico D é modificado ao longo do tempo conforme a dinâmica dos pedestres, onde cada pedestre deixa um rastro por onde passa. O campo de piso escolhido para ser utilizado neste trabalho é o Dinâmico.

Dado que a probabilidade de transição é proporcional ao campo de piso dinâmico, se torna mais atrativo seguir os rastros de outros pedestres. O campo de piso dinâmico também é sujeito a um desaparecimento progressivo dos rastros.

O ambiente escolhido para realizar os testes de evacuação é uma representação da planta do restaurante universitário da Universidade Federal de Catalão, como pode ser observado na Figura 3. O ambiente possui dimensões de 23,5m x 18,25m, que ao serem adaptados para o Látice do autômato passaram a ter 58 células (23,2m) x 40 células (18m), a saída possui a largura de 4 células (1,6m). O ambiente possui um total de 73 mesas (cada uma possui, em teoria, 4 cadeiras, o que não será considerado na simulação), com dimensões de 4 células (1,6m) x 2 células (0,8m), onde elas são organizadas de forma individual, ou em dupla, formando, assim, uma mesa estendida.

5. Testes

A partir do ambiente descrito anteriormente, foram selecionadas 5 possíveis lotações, sendo elas 292 pedestres (100% das cadeiras ocupadas), 219 pedestres (75% das cadeiras ocupadas), 146 pedestres (50% das cadeiras ocupadas), 73 pedestres (25% das cadeiras ocupadas), por fim também foi selecionada a lotação de 304 pedestres, que é a capacidade

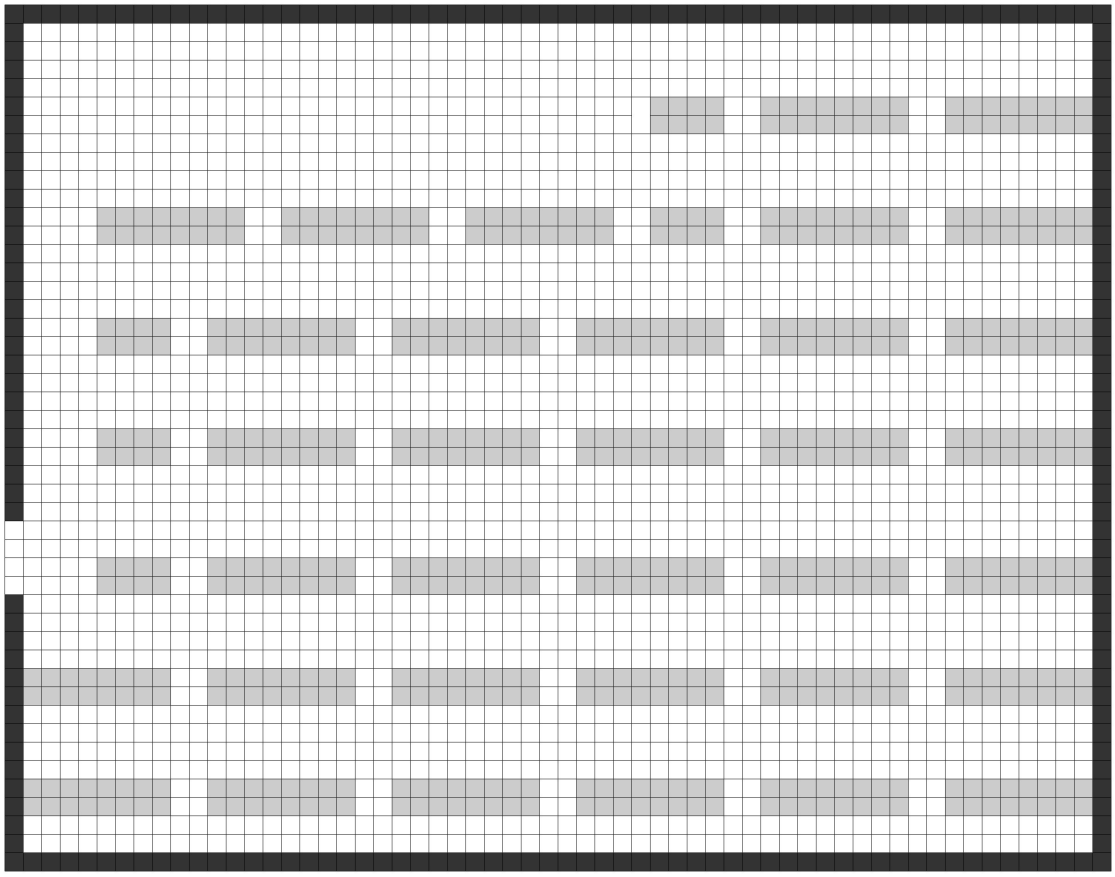


Figura 3. Representação de um Látice

máxima do ambiente, desconsiderando a quantidade de pedestres que podem, necessariamente, serem acomodados nas mesas.

Para cada uma das lotações citadas, foram realizados 100 testes, resultando em um total de 500 testes, como parâmetro de avaliação foram considerados o tempo de evacuação e a quantidade de colisões que ocorreram durante o processo de evacuação.

6. Resultados

A Tabela 1 apresenta os resultados obtidos para cada uma das lotações supracitadas, começando em 73 e terminando em 304. Os valores presentes nos campos Tempo Médio, Média de colisões e seus respectivos desvios padrões, são obtidos a partir da média dos tempos e colisões obtidos a partir das 100 execuções realizadas para cada uma das diferentes lotações.

Como pode-se observar, quanto maior a lotação, maior é o tempo necessário para finalizar a evacuação, bem como a quantidade de colisões aumenta de acordo com o aumento da lotação do ambiente. Vale ressaltar que esse aumento não é diretamente proporcional, como pode ser notado ao comparar os tempos e colisões do teste realizado com 73 pedestres com o teste realizado com 146 pedestres; o segundo possui exatamente o dobro da quantidade de pedestres, porém o tempo de evacuação aumentou pouco mais de 57%, enquanto a quantidade de colisões mais do que quadruplicou, ao comparar o primeiro com segundo exemplo.

Pedestres	Tempo Médio	Desvio Padrão	Média de colisões	Desvio Padrão
73	45,50	3,71	1214,75	221,89
146	71,80	4,73	6364,31	601,07
219	100,77	5,41	16410,44	1011,36
292	128,33	6,19	30818,04	1385,15
304	133,09	5,85	33757,32	1431,66

Tabela 1. Resultados dos testes

A Tabela 2 representa o aumento percentual (denotado por a.p.) que ocorre nos três campos citados anteriormente, isto é, dos pedestres, do tempo médio e da média de colisões. O aumento percentual de um cenário é dado em termos do cenário anterior. Nota-se que em nenhum dos cenários avaliados houve um aumento proporcional entre a lotação do ambiente, o tempo médio de evacuação e a quantidade média de colisões.

Pedestres		Tempo Médio		Média de colisões	
Quantidade	a.p.	Valor (seg)	a.p.	Quantidade	a.p.
73	0	45,50	0	1214,75	0
146	100,00	71,80	57,81	6364,31	423,92
219	50,00	100,77	40,35	16410,44	157,85
292	33,33	128,33	27,35	30818,04	87,80
304	4,11	133,09	3,71	33757,32	9,54

Tabela 2. Aumento percentual

7. Conclusão

Neste artigo, foi apresentado um autômato celular que simula a evacuação de pedestres, inspirado no modelo apresentado por [Burstedde et al. 2001], o cenário que foi selecionado para simular a evacuação é uma representação do restaurante universitário da Universidade Federal de Catalão, dentro deste cenário foram realizados testes com 5 diferentes lotações, 73, 146, 219, 292 e 304 pedestres.

Dentre os resultados obtidos, nota-se uma certa similaridade aos tempo apresentados por [Papinigis et al. 2010], nele é descrito que, para uma porta com 1,5m de largura, a vazão é de 2,175 pessoas por segundo, o que para um cenário com 73 pessoas, resulta em um tempo de aproximadamente 33,56s, enquanto na simulação o tempo médio é de 45,5s, já para a lotação de 146 pedestres, o tempo idealizado pelo artigo é 67,13s, enquanto o da simulação é de 71,8s, o tempo se torna praticamente idêntico no cenário com 219 pedestres, sendo de 100,69s o tempo idealizado pelo artigo, e de 100,77s o tempo médio da simulação, no cenário com 292 pedestres o tempo idealizado é de 134,25s enquanto o da simulação é de 128,33s, por fim, com a lotação 304 pedestres o tempo idealizado é de 139,77s, e o tempo médio da simulação é de 133,09s.

Por fim, pode-se observar uma proximidade considerável entre os tempos obtidos na simulação, e os calculados a partir do artigo de [Papinigis et al. 2010], demonstrando assim uma eficácia interessante na simulação realizada neste trabalho, especialmente em ambientes com maior quantidade de pedestres.

Referências

Alizadeh, R. (2011). A dynamic cellular automaton model for evacuation process with obstacles. *Safety Science*, 49(2):315–323.

Alves, E., da Silva e Rosa, G., and Marra, R. (2011). Êxodo e sua contribuição à urbanização de 1950 a 2010. *Revista de Política Agrícola*, (2):80–88.

Burstedde, C., Klauck, K., Schadschneider, A., and Zittartz, J. (2001). Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A: Statistical Mechanics and its Applications*, 295(3):507–525.

de Oliveira Carneiro, L. (2012). Simulação de evacuação de multidão por autômato celular estudo de caso em um estádio de futebol.

de Souza Rosa, R. (2018). Simulação de evacuação de pedestres baseada em autômato celular. Universidade Federal de Goiás (UFG). Monografia de Graduação, 57 p.

Feliciani, C., Zuriguel, I., Garcimartín, A., Maza, D., and Nishinari, K. (2020). Systematic experimental investigation of the obstacle effect during non-competitive and extremely competitive evacuations. *Scientific Reports*, 10(1):15947.

Helbing, D. and Johansson, A. (2009). *Pedestrian, Crowd and Evacuation Dynamics*, pages 6476–6495. Springer New York, New York, NY.

Hoogendoorn, S. P. and Daamen, W. (2005). Pedestrian behavior at bottlenecks. *Transportation Science*, 39(2):147–159.

Lima, E. B. (2007). Modelos microscópicos para simulação do tráfego baseados em autômatos celulares.

Nations, U. (2019). World population prospects 2019. : Highlights (ST/ESA/SER.A/423), Department of Economic and Social Affairs, Population Division.

Papinigis, V., Geda, E., and Lukošius, K. (2010). Design of people evacuation from rooms and buildings. *JOURNAL OF CIVIL ENGINEERING AND MANAGEMENT*, 16:131–139.

Schadschneider, A., Klingsch, W., Klüpfel, H., Kretz, T., Rogsch, C., and Seyfried, A. (2011). *Evacuation Dynamics: Empirical Results, Modeling and Applications*, pages 517–550. Springer New York, New York, NY.

Schiff, J. (2007). Cellular automata: A discrete view of the world. *Cellular Automata: A Discrete View of the World*.

Tanenbaum, A. (2003). *Redes de computadores*. Elsevier.

Varas, A., Cornejo, M., Mainemer, D., Toledo, B., Rogan, J., Muñoz, V., and Valdivia, J. (2007). Cellular automaton model for evacuation process with obstacles. *Physica A-statistical Mechanics and Its Applications*, 382:631–642.

Velasquez, W. and Alvarez-Alvarado, M. S. (2021). Outdoors evacuation routes algorithm using cellular automata and graph theory for uphills and downhills. *Sustainability*, 13(9).

Wolfram, S. (1982). Cellular automata as simple self-organizing systems. Thechnical report, Physics Department, California Institute of Technology, Pasadena CA 91125.

Estudo de similaridade textual entre objetos de convênios do Ministério da Agricultura, Pecuária e Abastecimento no estado de Goiás

Douglas Farias Cordeiro[1], Leandro Rodrigues da Silva Souza[2], Renata Moreira Limiro[1], Núbia Rosa Da Silva[3]

[1]Universidade Federal do Goiás (UFG)
Campus Samambaia – Goiânia – GO – Brazil

[2]Instituto Federal Goiano (IFGoiano)
Campus Rio Verde – Rio Verde – GO – Brazil

[3]Universidade Federal de Catalão (UFCAT)
Catalão – GO – Brazil

cordeiro@ufg.br, leandro.souza@ifgoiano.edu.br

renatamlimiro@ufg.br, nubia@ufcat.edu.br

Abstract. *This paper proposes to carry out a study based on knowledge graphs usage to identify similarities between public agreements related with the brazilian Ministry of Agriculture, Livestock, and Supply, in the context of the state of Goiás, Brazil. Results reveal the main thematics groups of agreements, enabling data argumentation, which further analysis could explore in more detail.*

Resumo. *Este artigo tem como proposta realizar um estudo com base na utilização de grafos de conhecimento para a identificação de pontos de similaridade entre convênios públicos vinculados ao Ministério da Agricultura, Pecuária e Abastecimento, no contexto do estado de Goiás. Os resultados revelam grupos temáticos principais de convênios, possibilitando a geração de características derivadas, as quais podem ser utilizadas em análises posteriores.*

1. Introdução

O Ministério da Agricultura, Pecuária e Abastecimento (MAPA) é órgão responsável, na autarquia federal, pelo encaminhamento de ações voltadas à gestão de políticas públicas no âmbito da promoção à agricultura e ao agronegócio, assim como nos aspectos que tangem ao estabelecimento de normas e regulações de serviços enquadrados neste setor [Brasil 2021a]. Neste sentido, o MAPA possui uma estrutura organizacional que atua em uma série de frentes, as quais consideram desde questões voltadas aos processos burocráticos da administração pública até a elaboração de estudos laboratoriais voltados à inovação na agricultura.

Todo o arcabouço organizacional e estrutural do MAPA se converge na promoção do desenvolvimento sustentável de toda cadeia produtiva agrícola brasileira [Brasil 2021a]. Grande parte das estratégias e da distribuição dos recursos públicos para

tais propósitos são encaminhados através de convênios e contratos de repasse, no sentido de uma distribuição de descentralização dos valores aplicados.

Os convênios, termos de parceira ou contratos de repasse podem ser descritos como acordos realizados entre a União e demais entes da Federação, como estados e municípios, ou ainda para com organizações não-governamentais, com o objetivo de transferir recursos financeiros a serem aplicados na realização de um objetivo comum [Brasil 2021b]. Neste contexto, se destaca a participação direta de duas partes: o cedente, enquanto responsável pelo repasse do recurso em questão, e o convenente, o qual recebe o referido recurso.

Neste cenário, é de interesse e de grande relevância compreender as relações intrínsecas entre os acordos de transferências financeiras públicas de forma a apoiar as estratégias de investimentos e de proposição de políticas públicas. No contexto deste estudo será considerado como objeto de exploração os convênios públicos, especificamente no âmbito do MAPA para o estado de Goiás.

Aquém das motivações decorrentes das necessidades de melhorias das estratégias de repasse, os atributos inerentes aos registros de convênios públicos, em sua estrutura de armazenamento, não possibilitam de forma direta uma relação explícita entre diferentes instâncias, o que acaba por se tornar um obstáculo na gestão dos convênios em relação a um atendimento equiforme e que contemple as demandas essenciais e prioritárias.

Uma das possibilidades para mitigar esse tipo de problema é através da aplicação de métodos para identificação de similaridades entre elementos textuais, os quais possibilitem um agrupamento das entradas por meio de classes. Para tanto, métodos de cálculo de similaridade textual podem ser explorados em conjunto com estratégias de agrupamento [Majumder et al. 2016].

É interessante observar que as pesquisas com propósito de identificação de similaridade entre elementos textuais tem um foco consideravelmente voltado para exploração de mecanismos que permitam relacionar diferentes elementos por meio de suas semelhanças semânticas e o alinhamento entre termos [Mihalcea et al. 2006, Yang et al. 2018, Xiong et al. 2020].

Neste contexto, este artigo tem como objetivo apresentar um estudo aplicado voltado à identificação de padrões em dados textuais provenientes de convênios públicos do MAPA para o estado de Goiás por meio da utilização de grafos de similitude e do método de classificação de Reinert [Reinert 2001].

2. Metodologia

Os aspectos metodológicos do presente estudo são orientados a partir do processo proposto por [Fayyad et al. 1996] denominado de Descoberta do Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD). O KDD é composto por cinco etapas: seleção, pré-processamento, transformação, mineração e interpretação. Entre as vantagens deste processo se destaca o fato de ser interativo e iterativo, permitindo, em caso de necessidade, intervenções por parte do analista, assim como retorno a etapas anteriores para melhoria e adequação dos métodos utilizados e dos resultados face aos objetivos. A Figura 1 apresenta o processo KDD em detalhes.

A fase de seleção se refere à definição do conjunto de dados face ao problema a ser

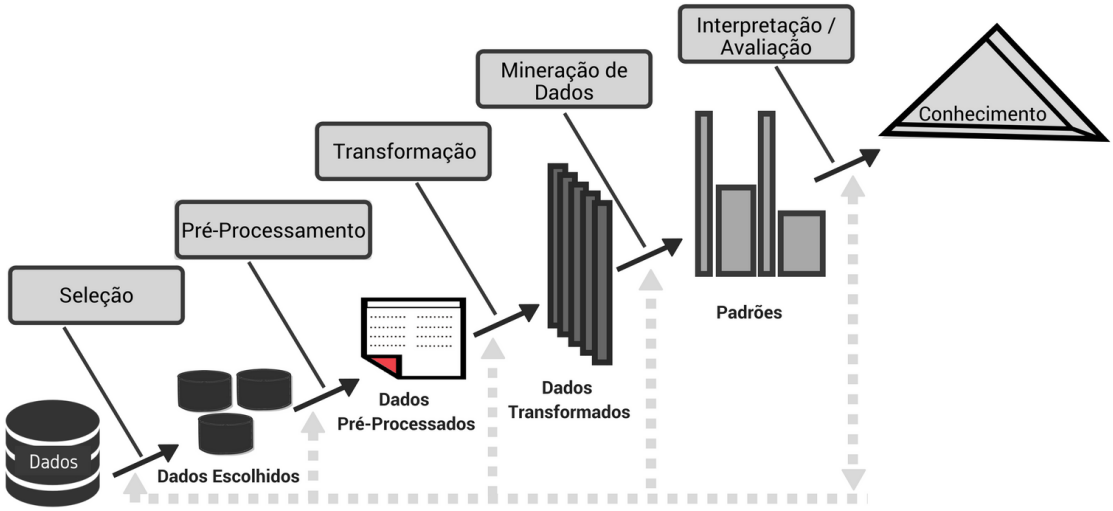


Figura 1. Processo KDD [Fayyad et al. 1996]

resolvido, assim como a obtenção dos referidos dados [Amaral 2016]. O problema central da presente proposta é realizar um levantamento das possíveis relações entre convênios públicos no estado de Goiás considerando o MAPA. Para tanto, na fase de seleção são utilizados dados abertos disponibilizados via Plataforma Mais Brasil[1]. A Figura 2 apresenta o diagrama entidade-relacionamento, sendo destacadas as duas entidades de interesse para as análises realizadas.

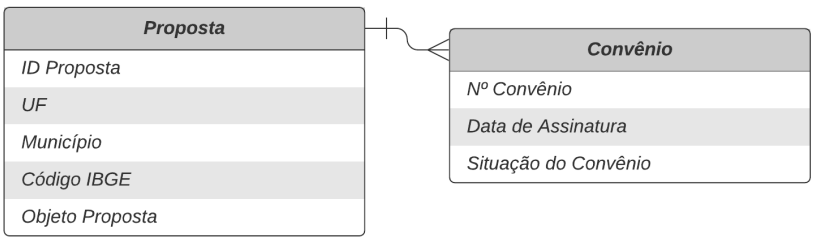


Figura 2. Diagrama Entidade-Relacionamento.

A etapa de pré-processamento trata da aplicação de soluções que possibilitem a garantia da qualidade os dados em termos de aplicabilidade das soluções de mineração de dados a serem utilizadas [Provost and Fawcett 2016] A partir da obtenção dos dados da Plataforma Mais Brasil, os mesmos foram sujeitos a rotinas de pré-processamento para filtragem ao subconjunto amostral de interesse, ou seja, especificamente dados para o estado de Goiás e com convênios ativos. Após isso, foram aplicadas rotinas de normalização textual, por meio de expressões regulares, sob o atributo ”Objeto Proposta”, de forma a remover inconsistências e instâncias vazias. A Tabela 1 apresenta algumas instância com a descrição dos objetos contemplados nos convênios.

[1]A Plataforma +Brasil se refere a uma solução integrada e centralizada, que disponibiliza dados abertos relacionados às transferências de recursos oriundos do Orçamento Fiscal e da Seguridade Social da União. Disponível em: https://portal.plataformamaisbrasil.gov.br/maisbrasil-portal-frontend/.

Id Proposta	Objeto
2213	Estamos propondo para o projetos de assentamentos no município de Formosa, construção de rede de distribuição elétrica, perfuração de vinte poços tubulares profundos, vinte reservatório de água 15000 litros, vinte casas de maquinas e 1000 m de tubo de pvc diâmetro de 60 mm, para instalação de um sistema de distribuição de água poço caixa d'água. A proposição deve-se á necessidade de suprir de água potável e com característica apropriada à agricultura familiar das famílias ali assentadas.
57157	O objeto do convênio é garantir recursos financeiros para a aquisição de máquinas e implementos agrícolas, equipamentos , ferramentas e insumos adequados para a execução das práticas de preparo do solo e cultivo das plantas medicinais, bem como construir um muro de proteção da área de plantio e um galpão para guarda de máquinas, implementos, ferramentas, insumos, etc. É importante também que nessa edificação tenha uma sala para a gestão da produção das plantas medicinais, onde será realizado o acompanhamento das atividades e o registro dos documentos específicos, obedecendo aos pops (procedimento operacional padrão) do setor, bem como a construção de banheiros para uso dos trabalhadores e uma copa para permitir que os mesmos guardem e preparem sua alimentação, possibilitando e oferecendo uma melhor condição de trabalho aos nossos colaboradores.
58342	Construção de ponte pré-moldada em concreto de 20,00x5,00 sobre o afluente do córrego Ribeirão Caldas.

Tabela 1. Exemplos de instância da base de dados.

Os dados pré-processados foram submetidos a uma etapa de transformação, de modo a estarem de acordo com o formato necessário para a solução utilizada, o software de mineração textual Iramuteq[2]. A etapa de transformação se refere à aplicação de rotinas para a adequação dos dados em termos estruturais, de forma a estarem em concordância com a solução algorítimica a ser utilizada [Grus 2021]. Neste sentido, foi gerado um corpus textual onde, para cada registro de convênio foram utilizados os campos de identificação e objeto de proposta.

A etapa de mineração de dados pode ser descrita como a aplicação efetiva de técnicas computacionais voltadas para a identificação de padrões, anomalias e possíveis correlações em grandes volumes de dados, de forma a gerar informações (resultados) [Dean 2014]. Neste sentido, a subárea de mineração de textos pode ser compreendida como aquela que tem foco específico na manipulação de dados textuais. Neste sentido, no presente estudo, por meio do uso do software Iramuteq foram aplicados os métodos de classificação de Reinert para obtenção das similaridades entre os elementos textuais [Reinert 2001], e a geração do grafo de similitude, o qual apresenta os principais termos e o relacionamento semântico entre os mesmos.

[2]http://www.iramuteq.org/

3. Resultados e Discussões

A Figura 3 apresenta o grafo de similitude obtido. Por meio deste é possível observar seis grupos principais de termos semanticamente interconectados, os quais vão no sentido de aquisição de equipamento, construção ou restauração de vias, produção agrícola, programas de apoio, desenvolvimento de projetos, sistemas de saúde animal e prevenção de pragas em plantas.

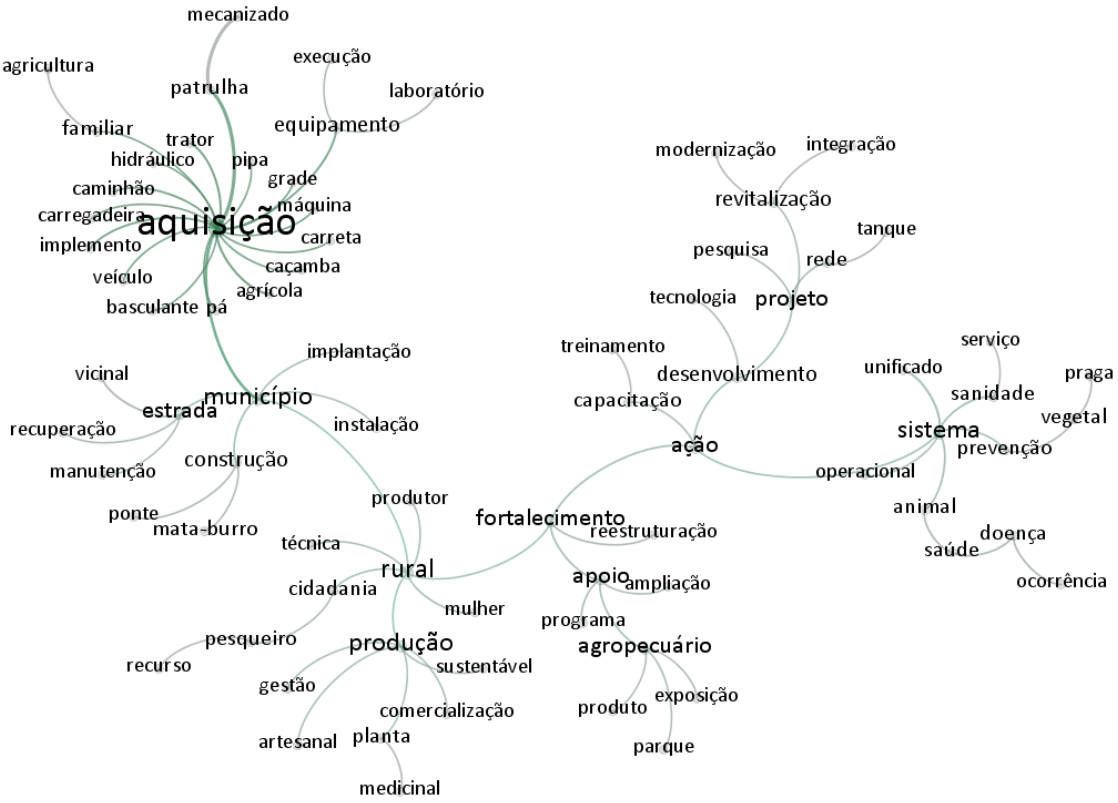


Figura 3. Grafo de similitude.

Por meio da aplicação do método de Classificação de Reinert foi possível se identificar as classes temáticas por meio da similaridade entre os elementos textuais. A Tabela 2, em detalhes, e a Figura 4 apresentam os termos de maior frequência para cada uma das classes, onde é possível observar a consonância com o grafo de similitude, se tendo uma consolidação de linhas de convênios para aquisição de itens, recuperação e construção, projetos de apoio ao agronegócio.

Finalmente, a Figura 5 apresenta os elementos textuais, no caso em questão os registros de cada um dos convênios analisados, com referência às classes identificadas. É interessante destacar que a partir desse resultado é possível realizar uma separação do conjunto amostral, de modo a explorar com maior profundidade detalhes e especificidades que possam ir de encontro ao apoio da promoção de políticas públicas no agronegócio no estado de Goiás, verificando, inclusive, questões quantitativas por meio de uma análise exploratória.

Classe	Termos
Classe 1	aquisição, patrulha, mecanizado, trator, agrícola, caminhão, grade, caçamba, carregadeira, pá, basculante, pipa, hidráulico, nivelador, pneu, pulverizador, distribuidor, calcário, arado, esteira
Classe 2	estrada, vicinal, ponte, mata-burro, recuperação, construção, instalação, farinha, fábrica, concreto, ração, contenção, assentamento, casa, readequação, manutenção, bueiro, peixe, unidade, obra
Classe 3	sistema, atenção, fortalecimento, apoio, ação, desenvolvimento, sanidade, produção, animal, saúde, doença, ocorrência, rede, revitalização, rural, defesa, mulher, sustentável, tecnologia, prevenção

Tabela 2. Vinte termos de maior frequência nas classes identificadas.

Figura 4. Dendograma de Classes.

4. Conclusão

Este artigo apresentou um estudo de aplicação de soluções de mineração de textos para a identificação de padrões em um conjunto amostral de dados de convênios públicos do MAPA para o estado de Goiás. Os resultados alcançados permitiram vislumbra as principais linhas explorados no âmbito de tais convênios, com destaque para a identificação de itens específicos. Por meio da identificação de classes, futuras análises com viés direcionado para o levantamento de questões quantitativas podem ser realizadas.

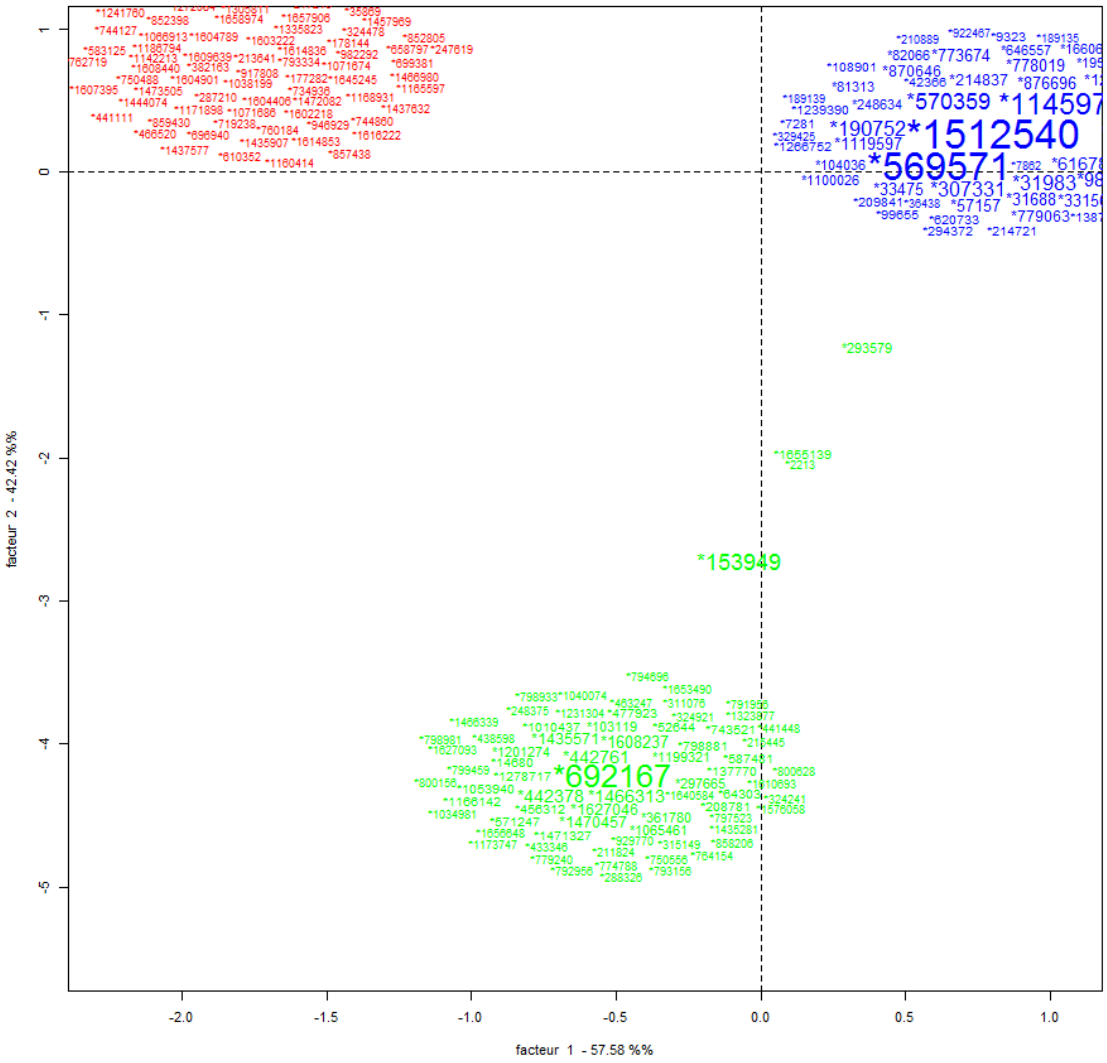


Figura 5. Identificação de elementos textuais por classe.

Referências

Amaral, F. (2016). *Aprenda mineração de dados: teoria e prática*. Alta Books, Rio de Janeiro.

Brasil (2021a). Ministério da Agricultura, Pecuária e Abastecimento. Institucional. Disponível em: https://www.gov.br/agricultura/pt-br/acesso-a-informacao/institucional. Acesso em 30 Out. 2021.

Brasil (2021b). Portal da Transparência. Convênios e outros acordos. Disponível em: https://www.portaltransparencia.gov.br/entenda-a-gestao-publica/convenios-e-outros-acordos. Acesso em 30 Out. 2021.

Dean, J. (2014). *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Wiley, Hoboken, Nova Jersey, EUA.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.

Grus, J. (2021). *Data Science Do Zero: Noções Fundamentais com Python*. Alta Books, Rio de Janeiro.

Majumder, G., Pakray, P., Gelbukh, A., and Pinto, D. (2016). Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas*, 20(4):647–665.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, page 775–780. AAAI Press.

Provost, F. and Fawcett, T. (2016). *Data science para negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados*. Alta Books, Rio de Janeiro.

Reinert, M. (2001). Alceste, une méthode statistique et sémiotique d'analyse de discours; application aux rêveries du promeneur solitaire. *La Revue Française de Psychiatrie et de Psychologie Médicale*, 5(39):32–36.

Xiong, Y., Chen, S., Qin, H., Cao, H., Shen, Y., Wang, X., Chen, Q., Yan, J., and Tang, B. (2020). Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. *BMC Medical Informatics and Decision Making*, 20(72).

Yang, Y., Yuan, S., Cer, D., Kong, S., Constant, N., Pilar, P., Ge, H., Sung, Y., Strope, B., and Kurzweil, R. (2018). Learning semantic textual similarity from conversations. *CoRR*, abs/1804.07754.

Análise da propagação da Covid-19 por meio de redes complexas

Erly de Araújo Lima Filho[1], **Douglas Farias Cordeiro**[2], **Núbia Rosa da Silva**[1]

[1]Instituto de Biotecnologia – Universidade Federal de Catalão (UFCAT)
Av. Dr. Lamartine Pinto de Avelar, 1200 Catalão – GO – Brazil

[2]Faculdade de Informação e Comunicação – Universidade Federal de Goiás
Goiânia – GO – Brazil

erlyalf@gmail.com, cordeiro@ufg.br, nubia@ufcat.edu.br

Abstract. *The pandemic caused by the corona virus, called Covid-19, has taken over the world in the last two years, causing millions of deaths and infected people. This pandemic further highlighted the need for studies related to the spread of diseases, as it showed how the world is unprepared to manage large spreads of diseases. Thus, this work aims to explore the spread of diseases through Covid-19 data modeling using complex networks.*

Resumo. *A pandemia causada pelo corona vírus, denominada de Covid-19, tomou conta do mundo todo nos últimos dois anos, tendo ocasionado milhões de mortes e pessoas infectadas. Esta pandemia evidenciou ainda mais a necessidade estudos relacionados a disseminação de doenças, pois mostrou como o mundo está despreparado para a gestão de grandes disseminações de doenças. Desta forma, este trabalho visa explorar a disseminação de doenças por meio da modelagem de dados da Covid-19 utilizando redes complexas.*

1. Introdução

O Sars-Cov-2, conhecido como coronavírus, vírus responsável pela Covid-19, se propagou em todo o mundo rapidamente, tendo início no final do ano de 2019. Desde então, mais de cinco milhões de óbitos no mundo foram confirmados e no Brasil, mais de seiscentos mil óbitos, devido à pandemia da Covid-19, de acordo com o Ministério da Saúde [Brasil 2021] tendo uma taxa de mortalidade de 86,5 a cada 100 mil habitantes.

Muitos estudos têm sido realizados utilizando redes complexas para simulação de diversas doenças, como por exemplo Dengue, Zica vírus, HIV, entre outras. No trabalho de Liu et al. [Liu et al. 2020] é analisado o surto do novo Corona vírus, onde é proposto um modelo epidêmico SAIR (suscetível - infectado assintomático - removido) para descrever o surto com base na epidemia de Wuhan. Para descrever o surto com precisão, foi estabelecido um modelo em uma rede social, que possui características de redes livre de escala, onde os nós representam os indivíduos e as arestas representam os contatos entre os indivíduos. Utilizando do cálculo de reprodução básica e de simulações de Monte Carlo, foi constatado que com o grau médio de ascensão da rede, a pandemia se espalharia mais rapidamente e o tempo de duração se tornaria mais longo. Portanto, constatam que outro método viável para conter a propagação da pandemia é reduzir a densidade das redes sociais, como limitar a mobilidade e os contatos sociais presenciais.

O trabalho [Stella et al. 2020] propõem o mesmo modelo epidêmico SAIR, mas com o foco no papel dos assintomáticos na pandemia. Para análise dos dados, diferentemente do trabalho anterior, foi utilizada a rede complexa pequeno mundo. Os estudos foram baseados no resultado que a abertura das escolas na Itália em setembro de 2020 pudesse causar. Foi descoberto que regiões da Itália com maior conectividade (conectividade essa que se dá por interação de uma região com outras regiões) teriam um aumento maior nos casos.

Zhan et al. [Zhan et al. 2020] investigam a dinâmica de propagação da epidemia COVID-19 por meio de otimizadores de reconhecimento. O modelo de rede complexas pequeno mundo é utilizada, onde os nós representam os indivíduos e as arestas são as ligações entre esses indivíduos. O modelo epidêmico SEIR (Suscetível – Exposto – Infectado – Removido) é utilizado para prever a propagação da epidemia em mais de 300 cidades na China. Porém, esse modelo tem mais de 1.800 parâmetros desconhecidos. O algoritmo de recozimento simulado pseudo-coevolucionário (SA), foi proposto para identificar esses parâmetros desconhecidos.

Resultados apontam que o número de infecções na maioria das cidades na China atingiu seu pico de 29 de fevereiro de 2020 a 15 de março de 2020. Para a maioria das cidades fora da província de Hubei, o número total de indivíduos infectados seria inferior a 100, enquanto para a maioria cidades na província de Hubei (excluindo Wuhan), o número total de indivíduos infectados seria inferior a 3.000. No trabalho de Sun et al. [Sun et al. 2020] é investigado como o COVID-19 impactou no transporte aéreo. No total, foram usados dados de serviços de 150 companhias aéreas entre 2.751 aeroportos. No total, foram usados dados de 152 dias, de 16 de dezembro de 2019 a 15 de maio de 2020. Foi analisado o sistema de aviação global como uma rede aeroportuária, com nós sendo aeroportos e voos que representam as conexões diretas entre aeroportos.

Além disso, em Sun et al. [Sun et al. 2020] também foi investigado a chamada rede de países, onde os nós são países e os links denotam a existência de voos diretos entre esses países. Foi constatado que na rede mundial de aeroportos que o hemisfério sul foi mais afetado do que a parte norte; ao considerar conectivo sozinho. A ligeira redução na distância entre os pares OD (origem e destino) mostrou que as restrições de voo foram impostas principalmente em voos internacionais de longa distância; consequentemente, os impactos da pandemia COVID-19 em voos internacionais foram muito mais fortes do que em voos domésticos. A análise das redes de países com foco no tráfego internacional de passageiros de 213 países no mundo mostrou que os padrões de conectividade para países individuais são heterogêneos e flutuam dependendo da situação do COVID-19.

São notáveis as contribuições alcançadas por meio da aplicação de redes complexas no contexto de construção de soluções que visem apoiar a compreensão, o combate e a prevenção de pandemias. Neste cenário, o objetivo deste estudo é modelar dados provenientes da epidemia da Covid-19 utilizando redes complexas com o intuito de mapear a disseminação da epidemia fornecendo subsídios para que sejam tomadas medidas de controle por meio de métodos de imunização. A teoria de redes complexas e métodos de propagação de epidemias serão usados para a modelagem de epidemias e obtenção do controle da doença.

2. Redes Complexas para modelagem do COVID-19

2.1. Redes livres de escala

Segundo [Barabási and Albert 1999] afirmam, algumas redes possuem uma organização na dinâmica de construção, com características particulares. Uma dessas características é quando um vértice tende a se conectar com outro vértice que possua muitas conexões, intitulada como conexão preferencial. Os *hubs* possuem essa característica, onde a rede possui poucos vértices altamente conectados e muitos com poucas conexões. A rede livre de escala é denominada assim pela sua forma matemática, ou seja, ela obedece uma função $f(x)$ que se mantém estável com um fator multiplicativo sob um re-escalonamento da variável independente x. Isto quer dizer que as redes livres de escala são aquelas que a distribuição de graus seguem a Lei de Potência, que consiste em um número pequeno de vértices com graus elevados e um número alto de vértices com graus baixos.

2.2. Modelo SIR

O modelo SIR (Suscetível-Infectado-Removido) criado por Kermack e McKendrick considera uma população fixa com três divisões: sensíveis S(t), infectado I(t) e removido R(t), que consistem em:

- S(t) representa o número de indivíduos não infectados com a doença no momento t, ou aqueles suscetíveis a doença.
- I(t) representa o número de indivíduos infectados com a doença e que são capazes de transmitir a doença para a categoria S(t).
- R(t) representa os indivíduos infectados que foram removidos da rede, ou a partir de uma morte ou de uma imunização. Essa categoria não é capaz de ser infectada novamente ou transmitir a doença para outras categorias.

2.3. Modelagem

Neste trabalho, uma previsão da pandemia em um determinado período de tempo é realizada, e comparada com os dados reais disponibilizados pelo Ministério da Saúde. Para extrair a taxa de transmissão, primeiramente foi utilizado a ferramenta analítica do COVID-19, disponibilizada pela Organização Pan-Americana da Saúde (PAHO), denominada EpiEstim [PAHO 2021]. O segundo passo foi aplicar essas taxas ao algoritmo de simulação em uma rede aleatória SIR, simulando um período de tempo de dez dias. O resultado dessa simulação é um gráfico contendo a proporção de infectados que é comparado ao gráfico de valores reais sem simulação.

Os dados utilizados sobre o COVID-19, foram extraídos da plataforma Brasil.io [Álvaro Justen 2021], que é alimentada pelos dados disponibilizados pelo Ministério da Saúde, mas de forma mais detalhada e em formato acessível. Para esse experimento, foram utilizados apenas dados da cidade de Catalão, Goiás. Foram trabalhados os dados do mês de Junho e Dezembro de 2020, e Junho de 2021, onde apenas as colunas *date* e *new-confirmed* foram utilizadas.

2.4. Base de dados

A base de dados utilizada do COVID-19, foi extraída do site Brasil.io [Álvaro Justen 2021]. Esse site é alimentado com informações vindas diretamente

do Ministério da Saúde, mas de forma acessível em arquivos *.csv*, com filtros variados. Esse formato favorece a manipulação dos dados pois todos os programas utilizados fazem uso de dados entrantes como arquivos *.csv* do Excel. Essa base contém diversas colunas, como semana epidemiológica, data, cidade, código do IBGE, novos casos confirmados e acumulados, entre outros. Mas para esse trabalho, apenas duas colunas foram utilizados, *date* e *new-confirmed*. Como o foco do trabalho foi apenas a cidade de Catalão, somente tais dados foram extraídos, fazendo uso do mês de junho e dezembro de 2020 e junho de 2021.

3. Resultados

Para início dos experimentos, o mês de Junho de 2020 é utilizado para extrair a taxa de transmissão. A ferramenta EpiEstim traz essa taxa de forma descomplicada e precisa, através de um arquivo *.csv* [PAHO 2021]. Este arquivo precisa apenas de duas colunas: *dates* e *"I"*, onde a primeira coluna representa cada dia de junho e a segunda coluna representa os infectados daquele dia respectivo, apenas. Então, para o mês de junho de 2020, a taxa de transmissão dada é de 0.99. Com a taxa de transmissão coletada, a mesma é inserida no código de simulação em um rede aleatória SIR, código este encontrado na biblioteca **EoN** do Python [Miller and Ting 2019].

Para o ambiente estar de acordo com Catalão, foram definidos como 111.000 o número de nós (população estimada em 2020 pelo IBGE). A partir deste momento, o ambiente de simulação está pronto. Para fazer um experimento fácil de visualização, foi escolhido um período de tempo de 10 dias para simulação, ou seja, esses dias são referentes aos primeiros dez dias de julho. O resultado dessa simulação está disposta na Figura 1.

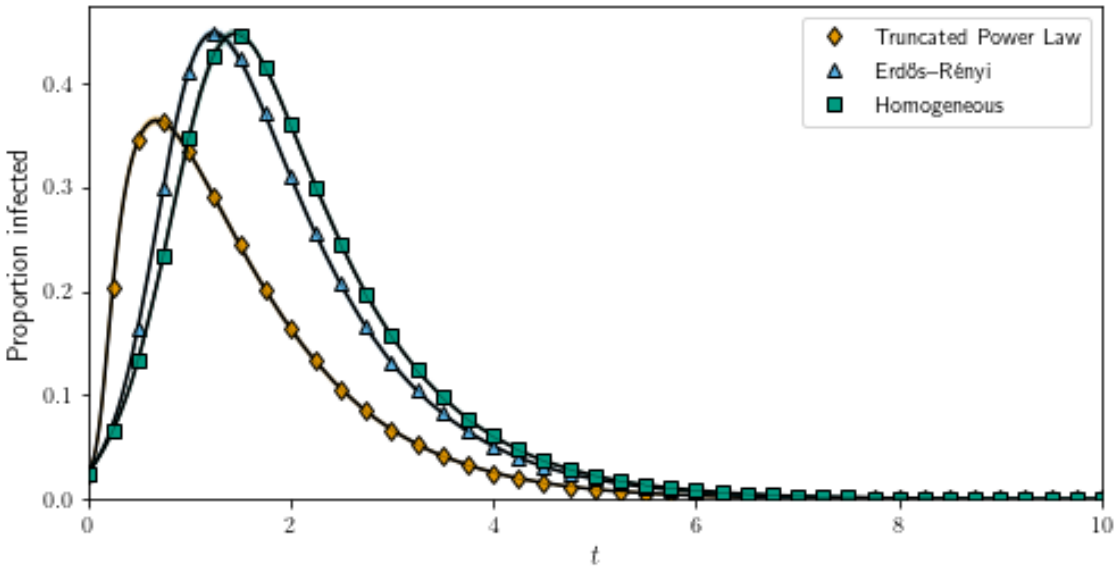


Figure 1. Simulação de 10 primeiros dias de julho de 2020. Fonte: Imagem elaborada pelo autor

Para comparar os dados simulados de junho com os reais, o primeiro passo é extrair a taxa de transmissão dos dados reais. Sendo assim, aqui é utilizado os dados não

apenas de junho, mas também os dados até o décimo dia de julho para se igualar ao tempo simulado anteriormente. Então temos que para esse período a taxa de transmissão é de 1,25. Como aqui não precisamos simular, e sim apenas visualizar a disposição da rede, o código é rodado sem o período de tempo de 10 dias, apenas mostrando o estado real momentâneo. O resultado dessa visualização está disposta na Figura 2.

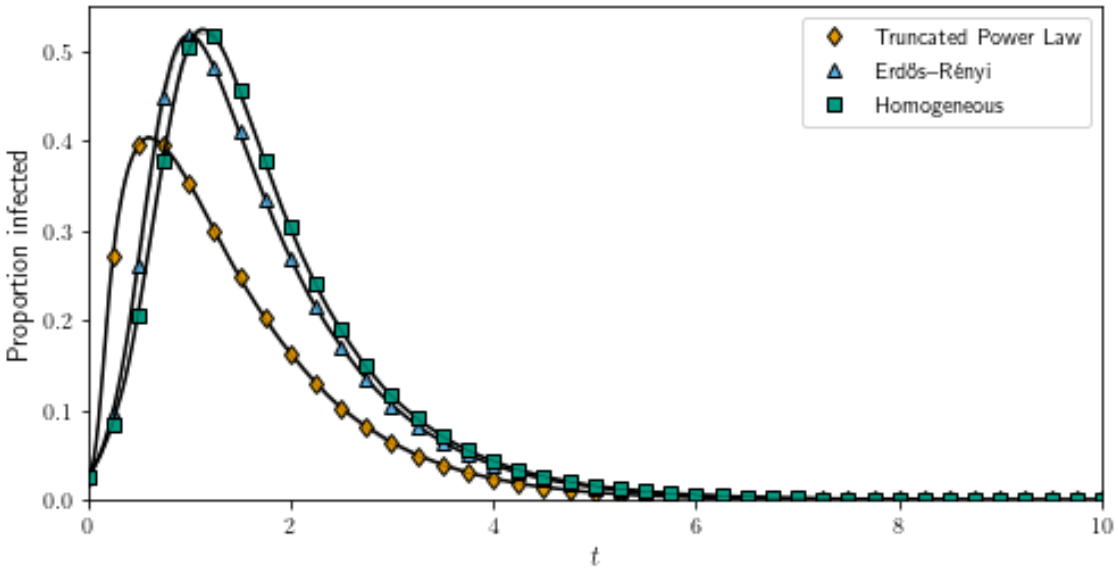


Figure 2. Dados reais junho/julho de 2020. Fonte: Imagem elaborada pelo autor

De maneira análoga, a simulação para dezembro de 2020 é realizada. Na Figura 3, temos a simulação de dezembro e na Figura 4, temos os dados reais. O mês de dezembro nos retorna uma taxa de transmissão de 1.05, que é utilizada para simular os primeiros 10 dias de janeiro de 2021. Utilizando-se dos dados reais até o décimo dia de janeiro de 2021 para se igualar aos dados simulados, temos uma taxa de transmissão de 1.62.

Do mesmo modo, a simulação de junho de 2021 é realizada. Na Figura 5, temos a simulação de junho de 2021 e na Figura 6, temos os dados reais. O mês de junho de 2021 nos retorna uma taxa de transmissão de 0.75, que é utilizada para simular os primeiros 10 dias de julho de 2021. Fazendo uso dos dados reais de junho até o décimo dia de julho de 2021 para se igualar aos dados simulados, temos uma taxa de transmissão de 1.1.

A partir dessas simulações, temos que o algoritmo utilizado da biblioteca **EoN** do Python, traz simulações muito próximas dos dados reais. Nota-se uma pequena diferença na proporção de infectados na rede aleatória de Erdös-Renyi comparando a simulação com os dados reais. Em julho de 2020, a simulação nos trouxe uma taxa de infectados acima de 0.4, enquanto que os dados reais nos trouxe uma taxa de infectados ligeiramente maior que 0.5. Da mesma forma, em janeiro de 2021, a simulação retorna uma taxa de infectados acima de 0.4 enquanto que para os dados reais, temos uma taxa de aproximadamente 0.6. E por fim, temos que nossa simulação de julho de 2021 nos traz uma taxa de infecção de aproximadamente 0.4, e em contrapartida temos que os dados reais possui uma taxa de aproximadamente 0.5. Ou seja, podemos afirmar que a proposta de simular a epidemia é válida e possui um resultado interessante que pode auxiliar em várias tomadas de decisões, favorecendo o controle de propagação da epidemia.

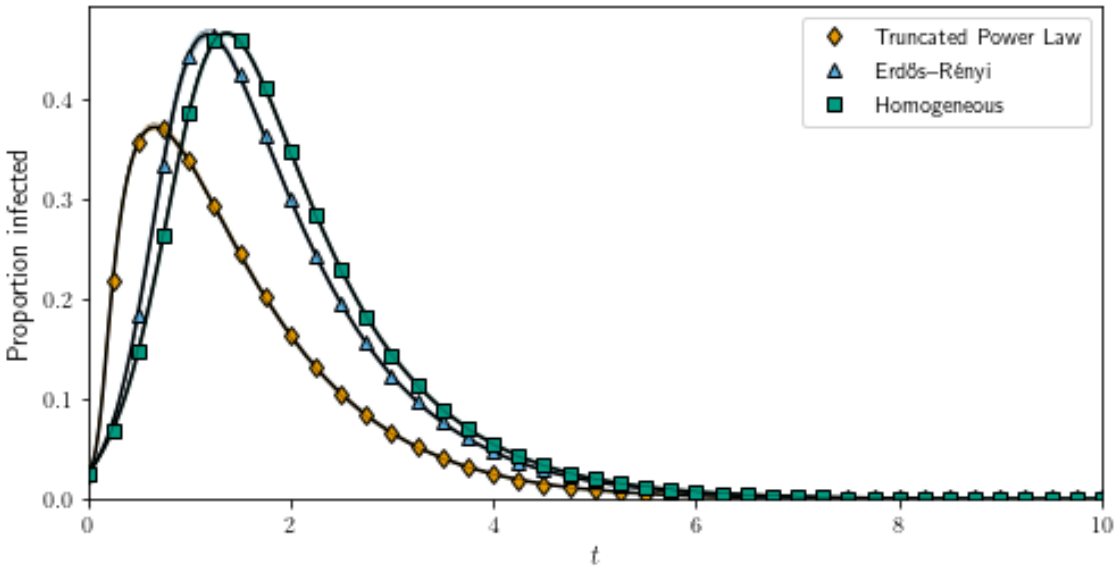


Figure 3. Simulação de 10 primeiros dias janeiro de 2021. Fonte: Imagem elaborada pelo autor

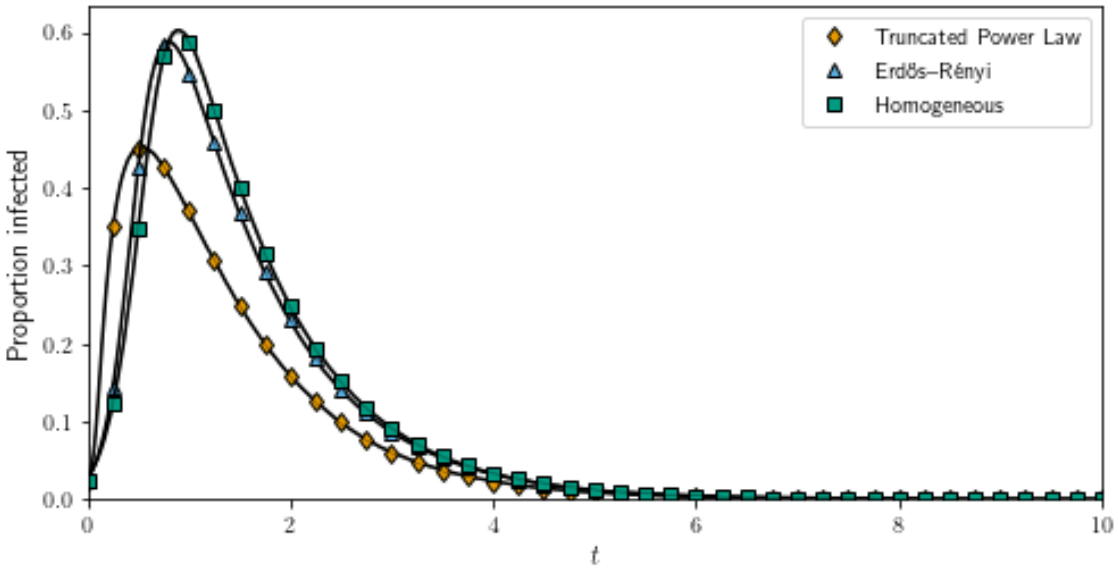


Figure 4. Dados reais dezembro/janeiro de 2021. Fonte: Imagem elaborada pelo autor

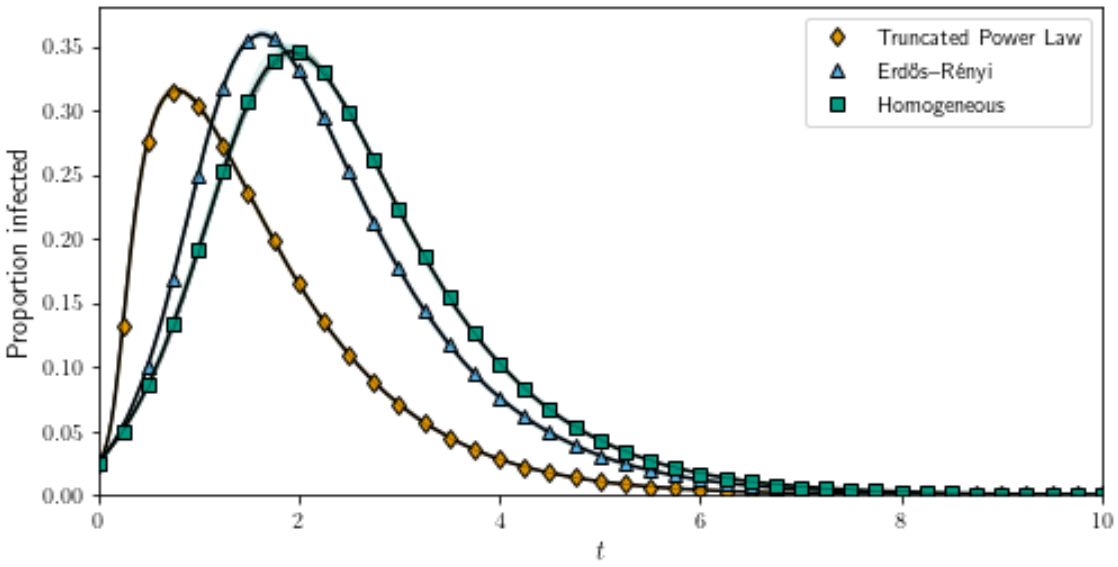


Figure 5. Simulação de 10 primeiros dias julho de 2021. Fonte: Imagem elaborada pelo autor

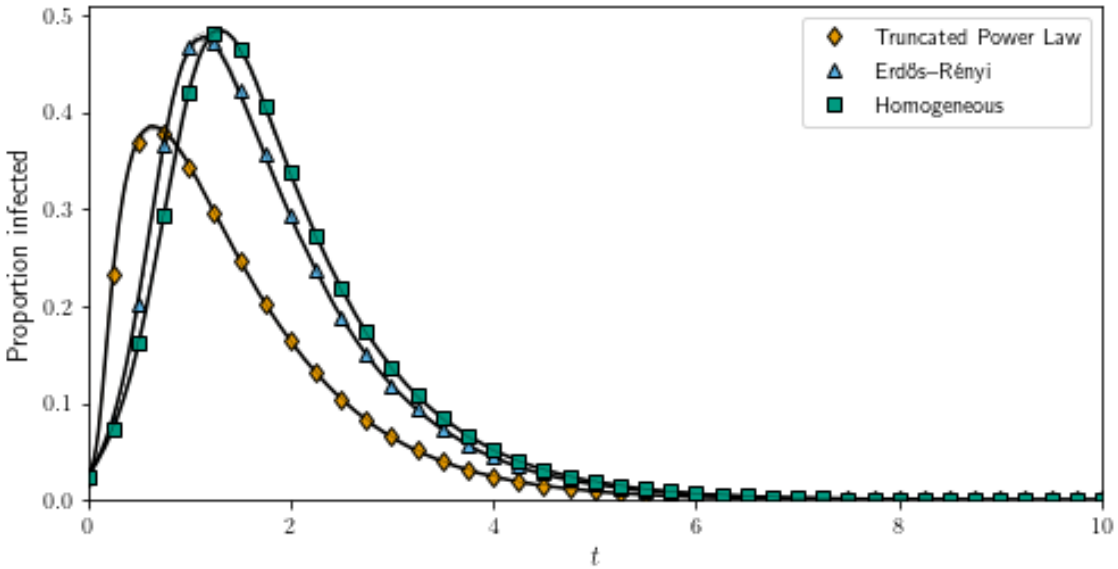


Figure 6. Dados reais junho/julho de 2021. Fonte: Imagem elaborada pelo autor

4. Conclusões

O desenvolvimento do presente estudo possibilitou uma análise de como a propagação de epidemias são realizadas. Além disso, também permite que tomadas de decisões sejam realizadas embasadas nos resultados da previsão da epidemia, decisões essas que afetam positivamente a propagação da epidemia, controlando a mesma. Ao fazer a simulação dos dados da Covid-19, foco central do trabalho, obteve-se números muito próximos dos dados reais, permitindo assim concluir que o código proposto é eficaz. Dada a importância do assunto, torna-se necessário o aperfeiçoamento das técnicas e estudo mais detalhado de ações a serem tomadas. Nesse sentido, para trabalhos futuros fica a validação de outros modelos epidêmicos e simulações mais detalhadas, para assim existir mais embasamento nas tomadas de decisões.

References

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Brasil (2021). Secretaria de vigilância em saúde - ministério da saúde. covid19 - painel coronavírus.

Liu, C., Wu, X., Niu, R., Wu, X., and Fan, R. (2020). A new sair model on complex networks for analysing the 2019 novel coronavirus (covid-19). *Nonlinear Dynamics*, 101.

Miller, J. C. and Ting, T. (2019). Eon (epidemics on networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. *Journal of Open Source Software*, 4(44):1731.

PAHO (2021). Covid-19 estimator. Disponível em https://harvardanalytics.shinyapps.io/covid19/. Acessado em 10/08/2021.

Stella, L., Martínez, A., Bauso, D., and Colaneri, P. (2020). The role of asymptomatic individuals in the covid-19 pandemic via complex networks. *Available at SSRN: https://ssrn.com/abstract=3688882 or http://dx.doi.org/10.2139/ssrn.3688882*.

Sun, X., Wandelt, S., and Zhang, A. (2020). How did covid-19 impact air transportation? a first peek through the lens of complex networks. *Journal of air transport management*, 89:101928.

Zhan, C., Zheng, Y., Lai, Z., Hao, T., and Li, B. (2020). Identifying epidemic spreading dynamics of covid-19 by pseudocoevolutionary simulated annealing optimizers. *Neural Computing and Applications*.

Álvaro Justen (2021). Repositório de dados públicos disponibilizados em formato acessível. Disponível em https://brasil.io/dataset/covid19/caso_full/. Acessado em 10/08/2021.

Redes complexas em dados sísmicos utilizando uma abordagem sequencial

Rafael Gomes Rodrigues[1], **Bruno Gomes**[1],
Douglas Farias Cordeiro[2], **Núbia Rosa da Silva**[1]

[1]Instituto de Biotecnologia – Universidade Federal de Catalão (UFCAT)
Av. Dr. Lamartine Pinto de Avelar, 1200 Catalão – GO – Brazil

[2]Faculdade de Informação e Comunicação – Universidade Federal de Goiás
Goiânia – GO – Brazil

rafael.computacao.gomes@gmail.com, brunog_100891@discente.ufcat.edu.br,

cordeiro@ufg.br,nubia@ufcat.edu.br

Abstract. *Complex networks have been used for several studies that use large amounts of data in a graph-shaped representation, where relevant information is obtained when studying the paths and structure of these graphs. This article presents an algorithm that uses seismic data from the time of occurrence of each earthquake to create the connections between them and thus form a complex network. For this, a seismic database is used, where a specific region is delimited and extracts data from earthquakes that happened in the period of five months in that region. As a result of the proposed algorithm, it was possible to observe some small world and non-scaled characteristics in the obtained network.*

Resumo. *As redes complexas têm sido utilizadas para diversos estudos que utilizam grande quantidade de dados em uma representação na forma de grafo, onde obtêm-se informações relevantes ao se estudar os caminhos e a estrutura desses grafos. Neste artigo apresenta-se um algoritmo que utiliza os dados sísmicos a partir do tempo de ocorrência de cada terremoto para criar as ligações entre eles e assim formar uma rede complexa. Utiliza-se para isso uma base de dados sísmicos, onde, delimita-se uma região específica e extrai dados de terremotos que aconteceram no período de cinco meses naquela região. Como resultado a partir do algoritmo proposto foi possível observar algumas características de pequeno mundo e sem escala na rede obtida.*

1. Introdução

Os terremotos têm sido estudados para se prever os possíveis locais de risco e as datas que podem acontecer estes abalos. Ao estudar estes dados foi possível constatar um padrão e representá-los através de uma rede. Estas redes são denominadas como redes complexas pois possuem extensa quantidade de informações (tempo, magnitude, espaço) que são relacionadas entre si e representadas em forma de grafo.

Apesar de já existirem estudos nesta área elas são concentradas em sua maioria nas regiões dos Estados Unidos e Japão [Abe and Suzuki 2004b, Abe and Suzuki 2004a, Abe and Suzuki 2007, Abe and Suzuki 2006], o que demonstra uma preocupação maior

desses países em analisar a ocorrência de terremotos. Para representar esses dados em redes complexas podemos utilizar dois parâmetros. O primeiro deles é a magnitude (Lei de Gutenberg-Richter [Gutenberg 2013]). Nesse tipo de trabalho são classificados os terremotos de acordo com o grau de magnitude, onde serão classificados cada terremoto como primários e secundários. Os terremotos primários são aqueles com os maiores valores de magnitude que representarão os terremotos principais "causadores" dos demais terremotos na região [Abe and Suzuki 2004b, Abe and Suzuki 2004a, Abe and Suzuki 2007, Abe and Suzuki 2006]. O segundo parâmetro é o tempo (Lei de Omori). Nesse tipo de abordagem, os terremotos principais são aqueles que aconteceram primeiramente naquela região, causando a ocorrência dos demais. Apesar dos dados do parâmetro de tempo serem mais consistentes, a maioria dos trabalhos encontrados utilizam a magnitude como principal aspecto para estudo utilizando redes complexas.

Estudos recentes na área comprovam que a representação de dados sísmicos através de redes complexas conseguem prever em curto prazo a ocorrência de terremotos, como aponta o estudo de [Papadopoulos 2016]. Outro trabalho a ser destacado utiliza redes bayesianas para estudar a predição de terremotos [Zhang et al. 2016], possibilitando a realização de previsão no intervalo temporal compreendido entre 01 de Janeiro de 1992 e 01 de Janeiro de 2012, com uma taxa de precisão de 65%. Em [Rezaei et al. 2017] é criado um modelo híbrido que constrói uma rede de terremotos a partir da extração das duas principais leis do campo dos terremotos supracitadas.

Neste artigo conduz-se uma representação dos dados sísmicos do estado do Alabama, localizado nos Estados Unidos da América, empregando o parâmetro do tempo de ocorrência dos abalos, em um período entre cinco meses. Para isso foi criado um algoritmo que capta o tempo dos acontecimentos dos terremotos e cria-se um relacionamento entre eles a partir de uma linha do tempo, assim é construída uma representação em rede complexa que contém características de redes de mundo pequeno[5] [Watts and Strogatz 1998] e sem escala[6] [Barabási and Albert 1999].

2. Revisão Bibliográfica

Em seu primeiro trabalho sobre redes complexas no ambiente sísmico, Abe e Suzuki [Abe and Suzuki 2004b] construíram um grafo representando os terremotos da região da Califórnia. Eles observaram que a rede apresentava características de uma rede livre de escala e também de pequeno mundo. Para um estudo mais aprofundado do tema os autores buscaram trazer as definições encontradas nos trabalhos de Watts e Strogatz, [Watts and Strogatz 1998] em redes de pequeno mundo[5] e de Barabási e Albert em redes livres de escala[6][Barabási and Albert 1999]. O conceito utilizado para representar a rede de terremoto é o seguinte[Abe and Suzuki 2004b]:

- Uma determinada região geográfica é dividida em células cúbicas;
- Esta célula é um vértice se ocorrerem terremotos com qualquer valor de magnitude;
- Se ocorrerem terremotos em diferentes células, estas serão ligadas através de uma aresta;

[5]As redes de pequeno mundo são caracterizadas por possuírem um grau de separação pequeno, ou seja, dado dois vértices aleatórios o caminho entre eles será sempre pequeno.

[6]As redes livres de escala consistem em redes complexas que são caracterizados por possuírem uma distribuição de conectividade que decai por uma lei de potência.

- Se acontecer terremotos sucessivos em uma mesma célula é colocado um loop no vértice.

Neste tipo de construção o único parâmetro é o tamanho da célula, em [Abe and Suzuki 2004b] foram utilizados os valores de 5km x 5km x 5km até 10km x 10km x 10km.

A partir da Figura 1, pode-se observar que a rede de terremotos possui alguns vértices conhecidos como "mainshocks" (vértices A e B), que geralmente ocasionam a distribuição de tensão sísmica, criando assim os outros vértices (terremotos secundários) [Abe and Suzuki 2004b]. Um fato identificado é que tremores secundários associados a esses "mainshocks" tendem retornar a vizinhança deste terremoto primário, o que consequentemente causa a incidência de "hubs" que possuem um alto grau de conectividade, o que leva o autor a pensar na possibilidade de ser livre de escala [Barabási and Albert 1999].

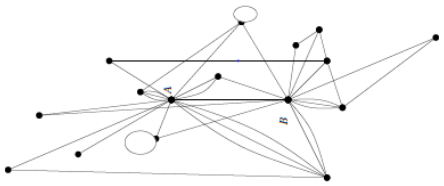


Figure 1. Grafo da rede de terremotos [Abe and Suzuki 2004b].

Na Figura 2 há a representação de graus de separação (comprimento do caminho) entre um par aleatório de vértices. Foram empregados diferentes tamanhos de célula e foram calculados em uma amostragem aleatória 60 pares de vértices. O que se notou é que os valores desse grau de separação foram pequenos variando entre 2 e 3, revelando assim a natureza do mundo pequeno.

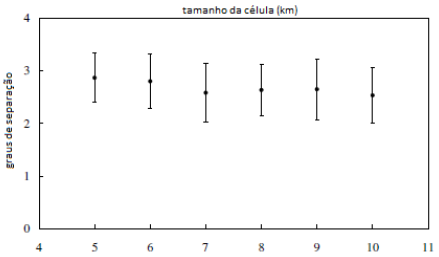


Figure 2. Graus de separação [Abe and Suzuki 2004b]

O estudo de Sumiyoshi Abe e Norikazu Suzuki [Abe and Suzuki 2004a] trata da definição de uma rede em evolução associada a terremotos, aplicando estes conceitos aos dados sísmicos recolhidos no sul da Califórnia e no Japão. Possui como ponto de destaque, e distinção de seus outros trabalhos, o objetivo de provar que as redes em evolução nessas duas áreas são de fato livres de escala. Uma rede livre de escala consiste em uma rede complexa com grau de distribuição seguindo a lei de potência, onde a maioria dos vértices possuem poucas ligações, em embate com a existência de alguns

vértices que apresentam um grande número de ligações, sendo que, um vértice de grau alto tende a ligar-se a outro vértice de grau alto.

Assim, a pesquisa é realizada na Califórnia e Japão, com regiões divididas em células cúbicas de 10km e 5km, formando vértices, em caso de terremoto, e ligando-se à outras áreas desenvolvendo arestas. De tal modo, que os dados sísmicos foram mapeados para um grafo aleatório. Por fim, descobre-se que as redes de terremotos possuem natureza livre de escala em suas distribuições de conectividade, portanto, apresenta-se um atributo novo do terremoto como um fenômeno crítico complexo. O experimento mostra que os tremores secundários associados à um tremor inicial tende a retornar à sua origem geograficamente, logo, contribui para o grande grau de conectividade do vértice do tremor inicial. Consistindo na origem da natureza da rede livre de escala do terremoto.

Para representação das distribuições de conectividades no sul da Califórnia inclui-se a Figura 3. Em (a) temos células de tamanho 10km x 10km x 10km, $y = 1,36$ e $k_0 = 1,65$. Em (b) a célula apresenta tamanho de 5km x 5km x 5km, $y = 1,61$ e $k_0 = 2,04$. Na Figura 4 há as distribuições de conectividades no Japão. Em (a) o tamanho da célula 10km x 10km x 10 km, $y = 2,22$ e $k_0 = 1,71$. Em (b) O tamanho da célula 5km x 5km x 5km, $y = 2,50$ e $k_0 = 0,79$. Na Figura 5 é representado a mudança do valor de k0 no sul da Califórnia de acordo com a evolução da rede sísmica. Nesta ocasião, o tamanho da célula agregada é 5km x 5km x 5km com $y = 1,61$.

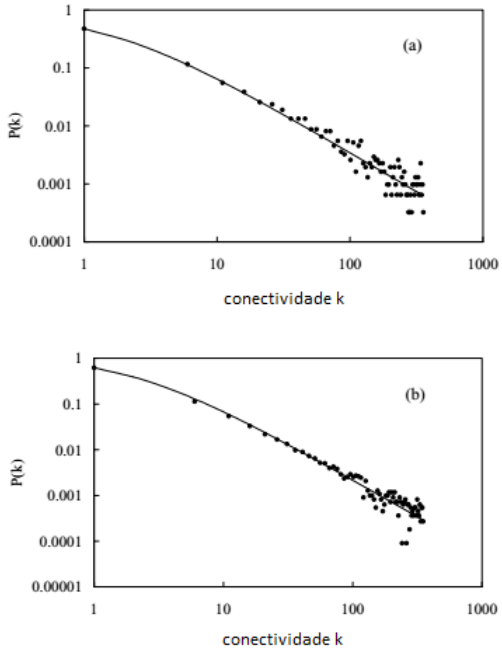


Figure 3. Conectividades no Sul da Califórnia. Fonte: [Abe and Suzuki 2004a]

Em [Abe and Suzuki 2007] é realizada uma discussão sobre a propriedade da rede complexa para terremotos construída sobre a Califórnia, construção citada em artigos anteriores [Abe and Suzuki 2004b, Abe and Suzuki 2004a]. A partir desta discussão é feito um experimento sobre a rede complexa, sendo descoberto que os valores do coeficiente de agrupamento da rede permanecem uniformes até sofrerem um ”salto” durante o choque/abalo sísmico inicial, então o valor cai gradativamente até um estado “morto”.

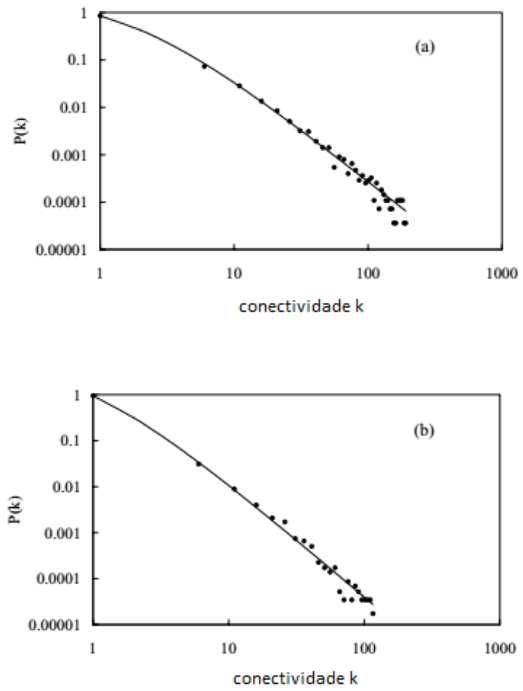


Figure 4. Conectividades no Japão. Fonte: [Abe and Suzuki 2004a].

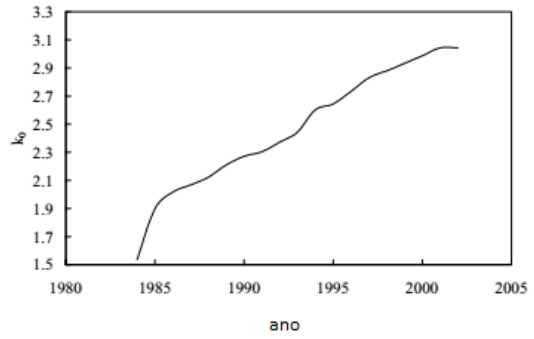


Figure 5. Mudança do valor de k0 no Sul da Califórnia. Fonte: [Abe and Suzuki 2004a].

Um abalo sísmico é medido por um conjunto de valores como o tempo de ocorrência, hipocentro e magnitude, onde o momento sísmico e a força de impacto são definidos em um espaço de tempo discreto. Diferente de outras áreas de estudo, ambos, momento sísmico e força de impacto apresentam aleatoriedade, e é justamente por isso que pensamos na rede complexa para representar e modelar o fenômeno, através da rede complexa é possível assumir problemas que aparentemente são impossíveis de se mapear os eventos e fazer com tenham uma organização mais consistente para realizar experimentos, estudos e coleta de dados.

Após o mapeamento das regiões e construção do grafo utilizando a regra de dois vértices aleatórios, obtém-se uma matriz que armazenará os vértices e a quantidade de arestas, após isso foi utilizado o coeficiente de agrupamento e analisada uma cadeia de

eventos dentro de uma linha temporal, realizando coletas de dados em cada momento e fechando com uma comparação. Nos resultados é notado que o coeficiente de agrupamento se mantém uniforme, cresce para um valor bastante significativo, e decai com o tempo até que se atinja um valor uniforme. O artigo realiza a discussão sobre os eventos sísmicos, apresenta o coeficiente de agrupamento, e então realiza o experimento para mostrar que o evento prova este aumento no coeficiente, o que contribui para trabalhos futuros onde se pode estudar sobre as consequências desse agrupamento da rede, que áreas foram mais atingidas, e como uma modelagem pode vir a contribuir para o estudo de terremotos, visto que é uma área com grande relevância pelos meios de comunicação, mas com poucas ferramentas de prevenção, visto a aleatoriedade do fenômeno.

Em [Abe and Suzuki 2006] foi analisado a estrutura da rede não direcionada, a estrutura hierárquica e a propriedade de mistura da rede de terremoto. Mostrou-se que o coeficiente de agrupamento de declínio como uma lei de potência no que diz respeito à conectividade, manifestando na organização da hierarquia. Este fato combinado com os resultados anteriores obtidos, implica na existência de expressivas semelhanças entre a rede de abalos sísmicos e os resultados obtidos na web. O próximo passo foi o estudo da propriedade de correlação da rede de terremotos, calculando a conectividade média do vizinho mais próximo a correlação do coeficiente Pearson (do inglês, *Pearson correlation coeficient*) que é a correlação de grau-grau dada por $\overline{k}_{nn}(k)$. Ele é usado quando o $\overline{k}_{nn}(k)$ é uma função negativa, que significa que os vértices tendem a se conectar com outros vértices que possuem menos conectividade na rede.

Sumiyoshi Abe e Norikazu Suzuki construíram as redes de terremotos na Califórnia e no Japão empregando dois tamanhos diferentes de células, 10km x 10km x 10km e 5km x 5km x 5km como citado em [Abe and Suzuki 2004b] [Abe and Suzuki 2004a]. Posteriormente, mapeou-se os gráficos aleatórios crescentes dois conjuntos de dados disponibilizados em (http://www.data.scec.org/) com dados de 1 de janeiro de 1984 até 31 de dezembro de 2004, e pelo Instituto Nacional de Pesquisa para a Terra Ciência e Prevenção de Desastres (http://www.hinet.bosai.go.jp/) com dados de 3 de junho de 2002 até 31 de março de 2005. O número total de eventos é 379728 e 382639, respectivamente.

Essas redes de terremotos são pequenas e livres de escala. Para investigar sua estrutura hierárquica, primeiro analisa-se o coeficiente de agrupamento como uma função da conectividade. Esta quantidade é dada como se segue. Considere: $c_i = \frac{2\mathrm{e}_i}{k_i(k_i-1)}$, onde e_i é dada por $\mathrm{e}_i = (A^3)_{ii}$, com a matriz de adjacência $A = (a_{ij})$ de um simples gráfico [isto é, $a_{ij} = 1$ se os vértices ij são ligados e $a_{ii} = 0$] e k_i é o valor da conectividade do i-ésimo vértice. Em seguida, o agrupamento coeficiente, $\overline{c}(k)$ é definido por:

$$\overline{c}(k) = (\frac{1}{[NP_{sg}(k)]}) \sum_{i=1}^{N} c_i \delta_{k_i k} \tag{1}$$

no qual P_{sg} representa a distribuição de conectividade do gráfico simples. Observa-se que, calculando o coeficiente de agrupamento, os *loops* têm de ser removidos e as arestas múltiplas são substituídas por arestas simples, a fim de reduzir a rede completa para o correspondente ao gráfico simples.

Na Tabela 1 são apresentados dados sobre a evolução temporal da rede na

Califórnia, nos quais o número de vértices é representado por (N), o número de arestas é (E), o coeficiente médio de agrupamento é ¡ c ¿ e o comprimento médio do caminho é ¡ l ¿. Nos referentes dados, apenas ¡ c ¿ é calculado após a remoção de loops e substituindo bordas múltiplas por arestas simples.

Ano	1984	1984-86	1984-94	1984-2004
N	1200	1956	2999	3913
E	18090	53993	208915	379726
<k>	30,152	55,209	139,32	194,08
<c>	0,388	0,476	0,577	0,635
<l>	2,69	2,63	2,53	2,52

Table 1. Evolução temporal da rede na Califórnia [Abe and Suzuki 2006].

Outra análise executada refere-se à vizinhança média do vizinho mais próximo. Considera-se a probabilidade condicional, $P(\frac{k'}{k})$, em que um vértice de conectividade K está ligado a um vértice de conectividade k'. Em seguida, a conectividade média de vizinho mais próximo de vértices com conectividade k é definido por:

$$\overline{k}_{nn}(k) = \sum_{k'} k' P(\frac{k'}{k}) \tag{2}$$

Em contraste com a análise do coeficiente de agrupamento, neste ponto loops e múltiplas bordas devem servir para descrever quantitativamente a sismicidade do evento. Em ambos os casos (Califórnia e Japão), os vértices com grandes valores de conectividade tendem a se ligar entre si.

O resultado é apresentado na Tabela 2. Os dados são obtidos a partir da análise da conectividade média do vizinho mais próximo, a correlação do coeficiente é positivo, confirmando que as redes de terremoto são ligadas umas com as outras.

	10Km x 10Km x 10Km	5Km x 5Km x 5Km
Califórnia	$r = 0,285$	$r = 0,268$
Japão	$r = 0,161$	$r = 0,156$

Table 2. Análise da conectividade média do vizinho mais próximo. Fonte: [Abe and Suzuki 2006].

3. Metodologia

No presente trabalho foi desenvolvido um algoritmo que relaciona eventos sísmicos de determinada região, onde considera-se o tempo de ocorrência de cada evento, ou seja, a medida em que há o crescimento do grafo e inserção de dados, os vértices correspondentes criam uma relação apenas se um evento aconteça em um espaço de tempo curto. Nos artigos consultados para o trabalho foi percebido que este tempo era de no máximo dois meses, ou seja, caso um evento ocorresse próximo a outro, mas o evento seguinte fosse dois meses a frente não seria considerada relação. Na base utilizada USGS (*Earthquake Hazards Program*), tem-se uma cadeia de eventos que se relacionam, porém se após um período de tempo além do considerável um novo evento aconteça, o mesmo não é considerado e não é acoplado a rede.

O algoritmo com finalidade de relacionar os eventos sísmicos é considerado orientado a eventos, ou seja, cada vértice é um evento em uma área, e o algoritmo serve para

relacionar esses eventos baseando-se em critérios. O primeiro passo é delimitar a região de ocorrência dos eventos, onde o operador deve determinar qual região fará o estudo dos eventos, levando em consideração que para continuar o algoritmo é necessário que eventos tenham ocorrido na região alvo. O segundo passo é identificar esses eventos e torna-los vértices de um grafo, ou seja, cada evento sísmico se torna um vértice de um grafo, mas sem qualquer relação entre eles no momento atual. O terceiro passo é fazer a leitura da base de dados daquela região, pois cada evento que foi identificado e transformado em vértice possui um local de ocorrência, data e hora, e essas informações serão cruciais para a construção do grafo. O último passo irá tratar a criação das arestas, primeiro é procurado o vértice mais antigo no grafo que será marcado como início, então observa-se os eventos que ocorreram nas proximidades no dias seguintes e estabelece-se arestas entre o inicial e os eventos que ocorreram mais tarde no mesmo dia ou no dia seguinte, após trabalhar-se com o vértice inicial é feito a varredura nos vértices seguintes e se baseando na base de dados são encontrados os eventos que ocorreram em seguida, estabelecendo novas ligações.

Para criação do grafo foi empregado o programa Gephi que possibilita a manipulação de grafos, a ferramenta é resumidamente dividida em visão geral, laboratório de dados e visualização. Primeiramente no laboratório de dados insere-se a base de dados, contendo as arestas e nós da rede que será criada no programa. A seguir, na Tabela 3, tem-se a base de dados de uma pequena região do Alabama.

Nodes/Id	Label	Time
M 1.6 - 12 km ESSE of Lake View - Alabama	12LakeViewAlabama	19/03/2004 - 07:00
M 2.0 - 11 km ESSE of Lake View - Alabama	11LakeViewAlabama	20/03/2004 - 19:22
M 2.8 - 11 km WNW of Montevallo - Alabama	11Montevallo	20/03/2004 - 10:40
M 1.8- 11 km ESSE of Lake View - Alabama	11LakeViewAlabama	20/03/2004 - 14:10
M 1.6 - 14 km WSW of Helena - Alabama	14Helena	26/03/2004 - 11:56
M 2.4 - 13 km WSW of Helena - Alabama	13Helena	20/03/2004 - 09:09
M 1.9 - 14 km WSW of Helena - Alabama	14Helena	02/04/2004 - 00:43
M 2.1 - 13 km WSW of Helena - Alabama	13Helena	04/07/2004 - 04:23
M 3.3 - Alabama	NLAlabama	09/05/2004 - 08:56
M 3.6 - Alabama	NLAlabama	19/08/2004 - 23:51
M 2.8 - 14 km SE of Brent - Alabama	14BrentAlabama	28/08/2004 - 05:06
M 2.6 - 8 km WSW of Alabaster - Alabama	8AlabasterAlabama	28/05/2004 - 01:20
M 1.7 - 14 km E of Woodstock - Alabama	14WoodstockAlabama	20/03/2004 - 07:10

Table 3. Base de dados Alabama.

Na coluna Nodes/Id tem-se os nós da rede, que foram todos os eventos coletados em uma região do Alabama, tendo informações que tornam cada evento único, ou seja, cada evento tem um nome/localização e esta informação é a que o torna único naquela região podendo ser usado como seu Id. Na coluna Label tem-se o nome fantasia que irá aparecer no grafo caso o usuário permita essa ação. Por fim na coluna Time temos as informações de ocorrência, ou seja, quando aconteceu levando em consideração até a hora/minuto.

Após realizar o estudo da base de dados e execução do algoritmo na região estabelece-se as ligações/arestas, sendo necessário informar ao programa Gephi através da guia Laboratório de Dados.

Foi percebido que o evento de magnitude 1.6 foi o primeiro a ocorrer na região, e em seguida cinco eventos aconteceram em um espaço de termo curto e próximos ao inicial, então estabelecemos ligações, e assim as ligações foram se formando. Na coluna

Source temos o início da ligação, e na coluna Target o destino, na coluna Type temos o tipo de ligação que neste caso consideramos não-direcionado, Weight será o peso que a ligação/aresta terá que foi padronizado com o valor 2.

4. Experimentos e discussões

Na site do USGS pegamos a foto da região que a base de dados retrata e rodamos o algoritmo em cima dessa região, o que será apresentado nas imagens a seguir, os círculos são os eventos que estão na base de dados.

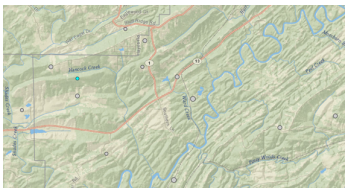


Figure 6. Trecho da região do Alabama [science for a changing world 2017].

Após rodar o algoritmo chegamos ao grafo apresentado na Figura 9.

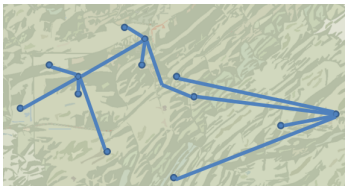


Figure 7. Grafo de eventos do Alabama.

No grafo conseguimos visualizar justamente a teoria das ondas sísmicas, onde determinado evento pode desencadear eventos ao redor nos dias seguintes, pois um choque inicial gera ondas que propagam na geografia da região, podendo gerar novos eventos, e estes podem gerar ainda mais ocorrências, que foi justamente o ocorreu no experimento, tivemos um choque inicial que seria o vértice de grau 5, e um dos vértices seguintes provocaram novos eventos e assim o grafo foi sendo formado. Outra questão marcante é a visualização dos *hubs*, que foram os vértices responsáveis por várias outras ocorrências ao redor dele, classificados como o epicentro da onda de choques sísmicos. O software então criou os nós e arestas, e através dos guias Visão Geral e Visualização tivemos a opção de ver o grafo pronto, e chegando a um grafo parecido com o do experimento feito em imagens da base de dados, onde foi obtido os mesmos resultados.

Figure 8. Grafo resultante do experimento.

Após feito o primeiro experimento com os dados do Alabama, os dados do USGS referentes a Califórnia foram inseridos, onde foi encontrado uma cadeia de 60 eventos que ocorreram entre os anos de 2014-2015 entre os meses de outubro – janeiro, então foi executado o algoritmo e feita a criação das relações.

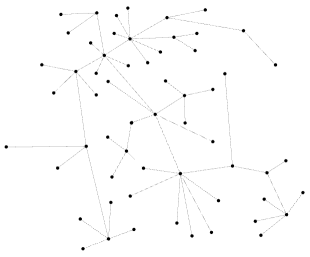


Figure 9. Grafo resultante - experimento Califórnia.

5. Conclusões

Os resultados obtidos neste trabalho demonstram que o parâmetro tempo utilizado para modelar a rede complexa se assemelha com os modelos que utilizam o parâmetro magnitude, pois os dois resultaram em redes com a característica de mundo pequeno e sem escala. Apresentando a incidência de nós hubs e também um caminho curto entre dois nós quaisquer da rede de terremoto.

References

Abe, S. and Suzuki, N. (2004a). Scale-free network of earthquakes. *EPL (Europhysics Letters)*, 65(4):581.

Abe, S. and Suzuki, N. (2004b). Small-world structure of earthquake network. *Physica A: Statistical Mechanics and its Applications*, 337(1):357–362.

Abe, S. and Suzuki, N. (2006). Complex earthquake networks: Hierarchical organization and assortative mixing. *Physical Review E*, 74(2):026113.

Abe, S. and Suzuki, N. (2007). Dynamical evolution of clustering in complex network of earthquakes. *The European Physical Journal B-Condensed Matter and Complex Systems*, 59(1):93–97.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Gutenberg, B. (2013). *Seismicity of the earth and associated phenomena*. Read Books Ltd.

Papadopoulos, G. A. (2016). Foreshocks and short-term hazard assessment of large earthquakes using complex networks: the case of the 2009 l'aquila earthquake. *Nonlinear Processes in Geophysics*, 23(4):241.

Rezaei, S., Darooneh, A. H., Lotfi, N., and Asaadi, N. (2017). The earthquakes network: Retrieving the empirical seismological laws. *Physica A: Statistical Mechanics and its Applications*, 471:80–87.

science for a changing world, U. (2017). Information by region-alabama.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442.

Zhang, Y., Zhao, H., He, X., Pei, F.-D., and Li, G.-G. (2016). Bayesian prediction of earthquake network based on space–time influence domain. *Physica A: Statistical Mechanics and its Applications*, 445:138–149.

Estudo estatístico da pandemia da Covid-19 no estado do Amapá

Douglas Farias Cordeiro[1], Renata Moreira Limiro[2], Núbia Rosa Da Silva[1]

[1]Universidade Federal do Goiás (UFG)
Campus Samambaia – Goiânia – GO – Brazil

[2]Universidade Federal de Catalão (UFCAT)
Catalão – GO – Brazil

cordeiro@ufg.br, renatamlimiro@ufg.br, nubia@ufcat.edu.br

Abstract. *This paper presents a study on the evolution of Covid-19 in the state of Amapá, Brazil, in terms of the number of notifications and death records. Therefore, the study is methodologically based on the Knowledge Discovery in Database (KDD) process, and uses measures of central tendency and dispersion to generate information. The results obtained present temporal and regional comparisons.*

Resumo. *O objetivo deste artigo é apresentar um estudo sobre a evolução da Covid-19 no estado do Amapá, em termos de número de notificações e registros de óbitos. Para tanto, o estudo se baseia metodologicamente no processo Descoberta do Conhecimento em Base de Dados (KDD), e utiliza de medidas de tendência central e de dispersão para geração de informação. Os resultados obtidos apresentam comparações temporais e regionais.*

1. Introdução

No final do ano de 2019 a China presenciou o surgimento de uma das pandemias mais alarmantes registradas pela humanidade, a pandemia da Covid-19, decorrente da Síndrome Respiratória Aguda Grave Coronavírus 2 (SARS-cov-2). Identificado como um coronavírus zoonótico, a Covid-19 ultrapassou na segunda metade do ano de 2021 o número de cinco milhões de óbitos ao redor do mundo. Campanhas vacinais têm sido realizadas com efetividade na maioria dos países. Apesar disso, é primordial conhecer os aspectos quantitativos relacionados à pandemia, de modo a apoiar o desenvolvimento de estudos e políticas públicas de saúde.

O presente artigo se refere a um estudo relacionado à aplicação de análise estatística para acompanhamento da evolução da Covid-19 no estado do Amapá. Para tanto são considerados dados públicos, disponibilizados pelo Ministério da Saúde, e enriquecidos pelo projeto Brasil.io. As análises realizadas apresentam indicadores que demonstram que, embora o estado do Amapá tenha uma densidade populacional relativamente menor que a de outros estados da região Norte, se figura, em determinados momentos, com valores preocupantes, principalmente em termos de mortalidade decorrente da Covid-19.

2. Procedimentos Metodológicos

A realização do presente estudo está baseada nos aspectos metodológicos descritos pelo processo conhecido como Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD) [Fayyad et al. 1996]. Embora o KDD seja voltado à aplicação de soluções baseadas em mineração de dados, o seu propósito principal é a geração de informação e descoberta de conhecimento. Neste sentido, compreende-se que o processo pode ser adaptado para procedimentos analíticos que se baseiam em técnicas e ferramentas provenientes de outras áreas ou campos de estudo, como é o caso da análise estatística.

O KDD é composto por cinco atividades sequenciais: seleção, pré-processamento, transformação, mineração de dados, interpretação [Dean 2014]. No contexto do presente estudo, que tem como foco analítico o uso de soluções estatísticas, a fase de mineração de dados não será realizada, sendo, neste caso, aplicadas as técnicas de medidas de tendência central e de dispersão [Spiegelhalter 2019]. A Figura 1 apresenta o esquema metodológico a ser utilizado.

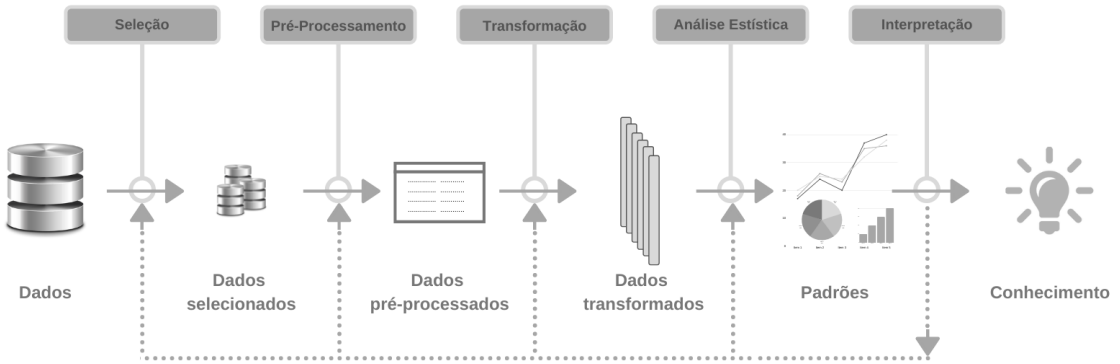


Figura 1. Processo KDD adaptado.

Para o presente estudo são considerados os dados de registros de novos casos de Covid-19, e de óbitos por Covid-19, para os estados da região Norte. Neste sentido, os dados foram obtidos por meio do projeto de disponibilização de dados públicos Brasil.io[1]. Apesar dos dados estarem diretamente disponibilizados pelo Ministério da Saúde, optou-se por utilizar essa fonte pelo fato de mesma disponibilizar a localização geográfica com os códigos do IBGE, facilitando a geração de análise georreferenciadas. O número total de registros presente na base é de 199.674. O conjunto de atributos considerado contempla:

- Código IBGE do estado;
- Código IBGE da cidade;
- Nome da cidade;
- Nome do estado;
- Data do registro;
- Número de novos casos;
- Número de novos óbitos;
- Estimativa de número de habitantes.

[1]http://brasil.io

Na etapa de pré-processamento de dados foi realizada a filtragem dos dados para os conjuntos amostrais a serem trabalhos, os quais contemplam dados de registros da região Norte, e para o período de Junho de 2020 e Junho de 2021. Além disso, ainda nesta etapa foram calculados os dados secundários: taxa de infecção (razão entre o número de notificação de casos e a quantidade de habitantes), e taxa de mortalidade (razão entre o número de óbitos e a quantidade de notificações de casos). Destaca-se que a base de dados não possui número de recuperados. Nos levantamentos realizados, tais dados também não foram encontrados no portal oficial do Ministério da Saúde.

Os dados obtidos e tratados foram armazenados de forma estruturada em arquivos no formato CSV, adequados para os propósitos de análise considerados no estudo, assim como para a manipulação através da ferramenta de visualização utilizada, o software de *self bi* Microsoft Power Bi[2].

Para realização da etapa de análise estatística foram consideradas as seguintes medidas de tendência central e dispersão:

- Média aritmética;
- Média móvel;
- Desvio padrão;
- Mediana;
- Moda;
- Quartis.

Além disso, são consideradas as gerações de gráficos comparativos. As medidas calculadas, assim como os gráficos obtidos, foram interpretados em face das comparações temporais possíveis entre ambos os conjuntos de dados, assim como comparações regionais.

3. Resultados e análises

A partir da aplicação do conjunto de rotinas para pré-processamento de dados, foram obtidos os conjuntos de dados referentes especificamente às análises a serem realizadas, as quais contemplam os dados de notificações de contaminação por Covid-19 nos estados da região Norte, assim como os registros de óbitos decorrentes desta doença para a mesma região. Neste sentido, se destaca que, remetendo-se ao objetivo do estudo, a proposta se refere à um estudo analítico, usando métodos estatísticos, prioritariamente nos meses de Junho de 2020 e Junho de 2021, para o estado do Amapá.

A partir disso, inicial foram calculadas as medidas estatísticas para os dados de número de notificações, conforme apresentado na Tabela 1. Juntamente à isso, foi calculada a evolução percentual entre os valores de Junho de 2020 e Junho de 2021. Inicialmente, percebe-se uma diferença considerável entre o número total de casos para os dois conjuntos amostrais, sendo a variação de aproximadamente -73,18%. De semelhante forma, todas as demais medidas apresentaram uma variação negativa considerável entre as amostras, salvo o valor mínimo, que na amostra de Junho/2020 foi de 84, e em Junho/2021 foi de 81.

Um dos pontos que chama a atenção nas medidas apresentadas na Tabela 1 se refere ao desvio padrão, o qual apresentou altos valores para ambos conjuntos amostrais.

[2]https://powerbi.microsoft.com/

Medidas	Junho/2020	Junho/2021	Evolução (%)
Total	18.890	5.066	-73,18%
Média Diária	629,67	168,87	-73,18%
Desvio Padrão	576,10	95,24	-83,47%
Moda	489,00	146,00	-70,14%
Mediana	239,00	184,00	-23,01%
25-Percentil	267,50	118,75	-55,61%
75-Percentil	671,50	184,00	-72,60%
Mínimo	84	81	-3,57%
Máximo	3.022	563	-81,37%

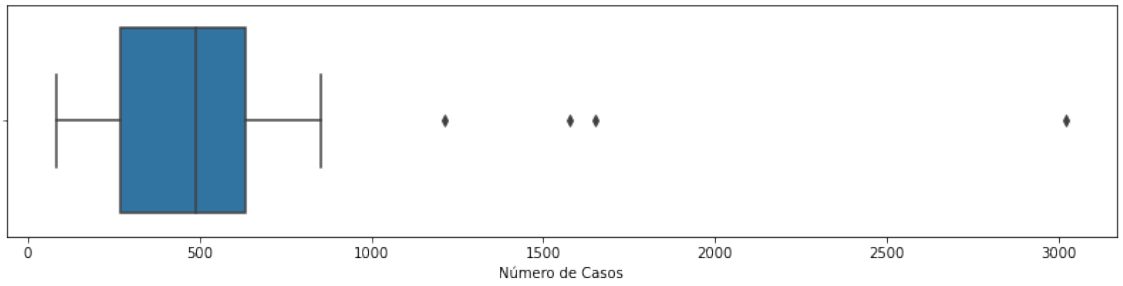
Tabela 1. Medidas estatísticas sobre o número de casos no estado do Amapá.

Isso significa que existe uma grande dispersão de dados, ou seja, os valores distam de forma significativa da média. Isso pode ser observado, de certo modo, ao se verificar as demais medidas. Para a moda, que se refere ao valor com maior frequência para os dados analisados, o conjunto amostral de Junho/2020 apresentou valor de 489, enquanto o conjunto amostral de Junho/2021 foi de 146. Ambos valores se afastam consideravelmente da média. Uma forma de explorar e compreender a dispersão dos dados, de maneira sumarizada, é através do *boxplot*.

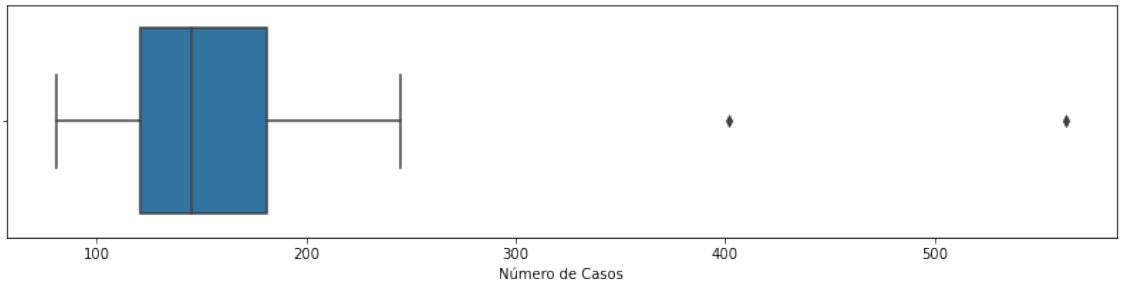
A Figura 2 apresenta o *boxplot* para o número de casos notificados em ambos conjuntos amostrais. É possível notar que na Figura 2-a, a linha central sobre o quadrado principal, a qual se refere à mediana, com valor 239, encontra-se ligeiramente à esquerda, o que significa que há uma dispersão maior de valores que estejam abaixo da mediana e acima do primeiro quartil (25-Percentil), que possui valor 267,50. Por outro lado, o conjunto de registros entre a mediana e o terceiro quartil (75-Percentil), com valor igual a 671,50, possui menor dispersão. Além disso, destaca-se ainda quatro registros com valores com alta dispersão em relação à média, os quais, teoricamente, são considerados *outliers*. Neste estudo, tais valores são importantes para a identificação dos picos de infecções notificadas, portanto não são sujeitos a tratamento.

Em relação à Figura 2-b, ao contrário do que ocorre na Figura 2-a, a mediana está mais deslocada em direção ao primeiro quartil, o que demonstra que há uma menor dispersão entre o conjunto de registros que se encontra nesta faixa. Por outro lado, para os dados entre a mediana e o terceiro quartil, há uma maior dispersão de dados. Para este caso, foram encontrados dois registros com valores, teoricamente, rotulados com *outliers*. De semelhante forma ao destacado anteriormente, tais valores são importantes para a identificação dos picos de registros presentes no conjunto amostral.

De semelhante maneira ao realizado com os registros de notificações de casos de infecção, foram calculadas as medidas de tendência central e dispersão para os registros de óbitos por Covid-19 para os conjuntos amostrais de dados de Junho de 2020 e Junho de 2021. A Tabela 2 apresenta em detalhes os valores calculados, assim como a evolução percentual em relação aos dois conjuntos amostrais. Uma característica que chama a atenção se refere à diferente entre as variações percentuais referentes ao número de casos notificados e de óbitos registros, que neste último apresente valores consideravelmente menores, ou seja, existe um indicativo que nos conjuntos amostrais como um todo, en-



(a) Junho de 2020.



(b) Junho de 2021.

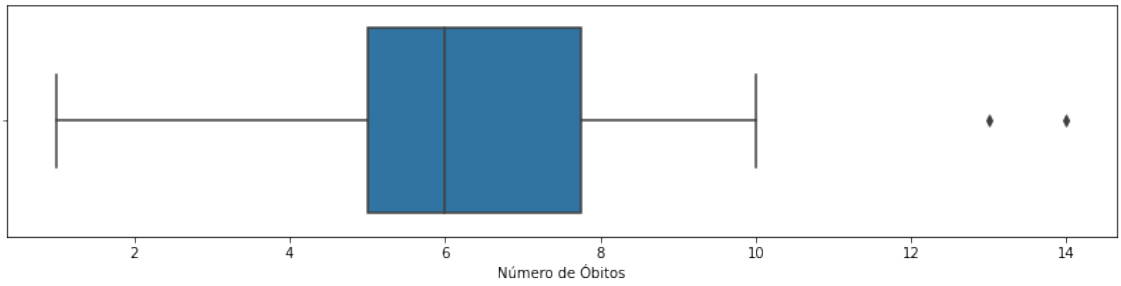
Figura 2. *Boxplot* para o número de casos dos conjuntos amostrais considerados.

quanto o número de casos diminuiu de Junho de 2020 para Junho de 2021, a quantidade de mortes proporcional ao número de casos aumentou. Isso pode ser visto se calculando a mortalidade para cada conjunto amostral, ou seja, a razão do total de óbitos pelo total de casos notificados, sendo a taxa de mortalidade igual a aproximadamente 1,03% para Junho/2020, e igual a aproximadamente 2,74% para Junho/2021.

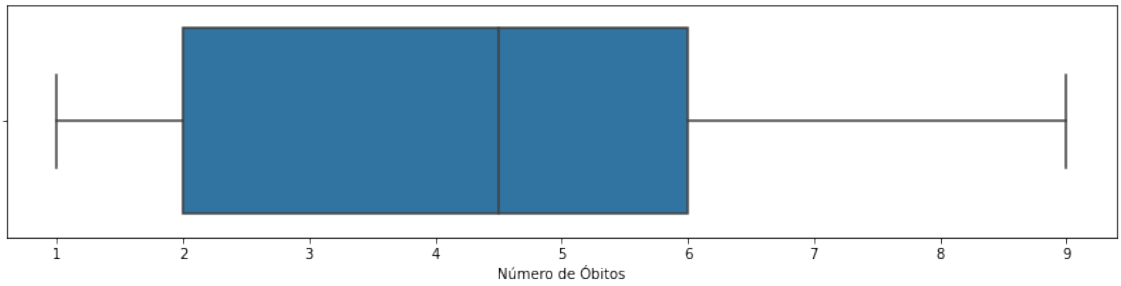
Medidas	**Junho/2020**	**Junho/2021**	**Evolução (%)**
Total	195	139	-28,72%
Média Diária	6,50	4,63	-28,72%
Desvio Padrão	2,81	2,48	-11,74%
Moda	7,00	6,00	-14,29%
Mediana	6,00	4,50	-25,00%
25-Percentil	5,00	2,00	-60,00%
75-Percentil	8,00	6,00	-25,00%
Mínimo	1,00	1,00	0,00%
Máximo	14,00	9,00	-35,71%

Tabela 2. Medidas estatísticas sobre o número de óbitos no estado do Amapá.

Ainda com relação às medidas apresentadas na Tabela 2, é possível notar que houve uma variação percentual de -28,72% da média, que caiu de 6,50 para o conjunto amostral de Junho/2020 para 4,63 para o conjunto amostral de Junho/2021. Entretanto, o desvio padrão, para ambos os casos, apresentou um valor relativamente alto, o que indica dispersão no conjunto de dados. De semelhante forma ao que foi realizado para os dados de notificação de casos, foram gerados os gráficos de *boxplot* para os registros de óbitos (Figura 3).



(a) Junho de 2020.



(b) Junho de 2021.

Figura 3. *Boxplot* para o registros de óbitos dos conjuntos amostrais considerados.

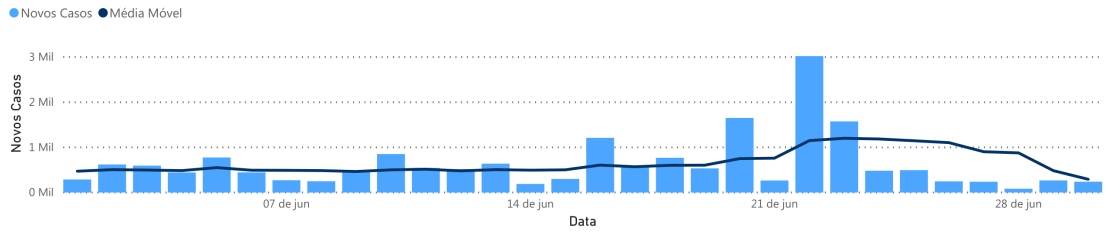
Na Figura 3, referente ao conjunto amostral de óbitos para Junho de 2020, é possível observar que a mediana, que possui valor igual a 6,00 está ligeiramente mais próxima do primeiro quartil, que tem valor igual a 5,00, ou seja, há uma dispersão proporcionalmente menor do que é observado entre a mediana e o terceiro quartil, que tem valor igual a 8,00. Além disso, é possível constatar que a dispersão entre o valor mínimo identificado, que é 1,00, e o primeiro quartil, é ainda maior. Além disso, o gráfico mostra a ocorrência de dois valores identificados como *outliers*.

Por outro lado, ao contrário do comportamento observado para o conjunto amostral de registros de óbitos de Junho de 2020, o conjunto amostral de Junho de 2021 apresenta uma dispersão maior entre primeiro quartil, com valor igual a 2,00, e a mediana, com valor igual a 4,50, e uma dispersão menor entre a mediana e o terceiro quartil, com valor igual a 6,00. Isso significa que 25% dos registros estão no intervalo [2,4], e outros 25% estão no intervalo [5,6]. Essa baixa dispersão entre a mediana e o terceiro quartil fica deveras evidente ao se verificar o valor da moda, que é igual a 6,00.

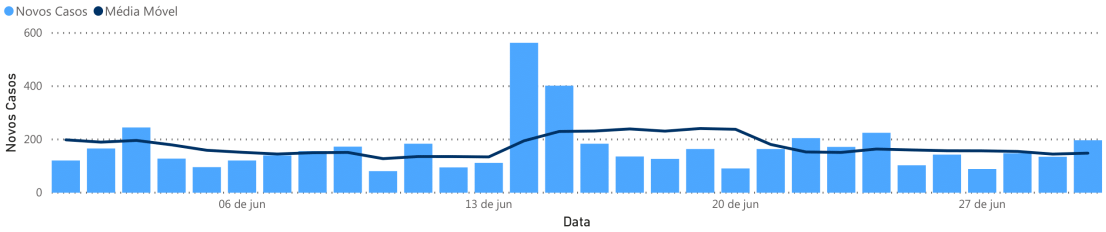
De modo a se compreender melhor, e em um maior nível de detalhamento, foram gerados os gráficos com as séries temporais para os conjuntos amostrais, tanto para as notificações de infecção, quanto para os registros de óbitos por Covid-19. A Figura 4 apresenta os gráficos em termos dos valores diários e da média móvel calculada, para um intervalo de sete dias. Na Figura 4-a, é possível notar que existe uma pequena variação dos valores na primeira metade dos registros, com um crescimento que segue até o pico, com valor igual 3.022. A média móvel, para este conjunto amostral, de semelhante forma, permanece com uma baixa variação nas datas que antecedem o dia 20 de Junho de 2020, quando existe a ocorrência de um pico, com valor de 1.652, alcançando seu máximo após o pico máximo, que ocorre em 22 de Junho de 2020, alcançando o valor máximo, igual a

1.201,86. Após isso, os últimos registros um decréscimo, sendo o valor mínimo da média móvel presente no último registro, em 30 de Junho de 2020, com valor igual a 293,57.

Para a Figura 4-b, referente aos casos notificados de Junho de 2021, é identificado um comportamento com uma variação relativamente menor que a verificada no conjunto amostral de Junho de 2020. A maior parte dos registros possui valor entre 100 e 200, com exceção dois registros consecutivos de valor 563 e 402. Esse comportamento pode ser observado através da média móvel, que possui valor mínimo em 127,86, e valor máximo em 239,43, resultante justante dos registros isolados de alto valor observados.



(a) Junho de 2020.



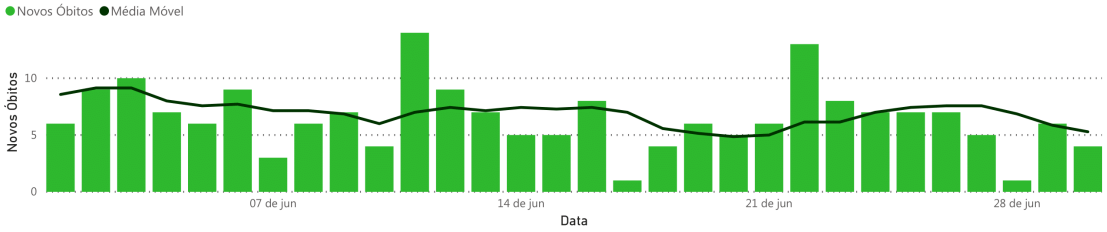
(b) Junho de 2021.

Figura 4. Evolução do número de casos notificados de Covid-19 no Amapá.

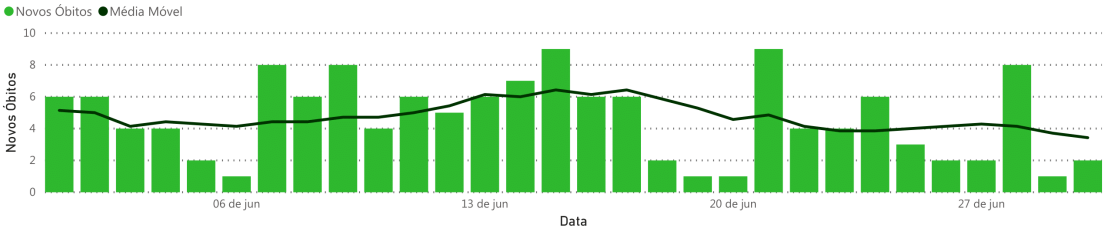
De maneira similar ao realizado com os dados de notificação de infecção, foram gerados os gráficos para os dados de óbitos por Covid-19, para os períodos temporais considerados no estudo, os quais são apresentados na Figura 5. Em relação ao conjunto amostral de Junho de 2020, é possível observar um comportamento de baixa variação, com alguns mínimos, os quais, em parte, estão relacionados aos finais de semana, onde ocorre uma menor oficialização dos registros. Tal comportamento é evidenciado pela média móvel dos últimos sete dias, que permanece entre seu valor mínimo de 4,86 e seu valor máximo de 9,14.

Para os dados do conjunto amostra de Junho de 2021, de semelhante forma, são constatadas pequenas variações, com mínimos relacionados essencialmente aos finais de semana. Esse comportamento remete aos valores de desvio padrão, apresentados na Tabela 2. A média móvel calculada apresentou valor mínimo de 3,43 e valor máximo de 6,43.

De modo a aumentar o nível de detalhamento e promover uma melhor compreensão, foram calculadas as taxa de contaminação e de mortalidade para os municípios do estado do Amapá, considerado como foco deste estudo. A Figura 6 apresenta os gráficos analíticos resultantes, incluindo um mapa de saturação, onde a cor mais clara se refere ao menor valor, e a cor mais forte, ao maior valor identificado, usando como referência o número de casos registrado para cada conjunto amostral. Nesta análise, em específico, é interessante observar, com relação ao mapa de saturação, que houve, de Junho de 2020



(a) Junho de 2020.



(b) Junho de 2021.

Figura 5. Evolução do número de óbitos por Covid-19 registrados no Amapá.

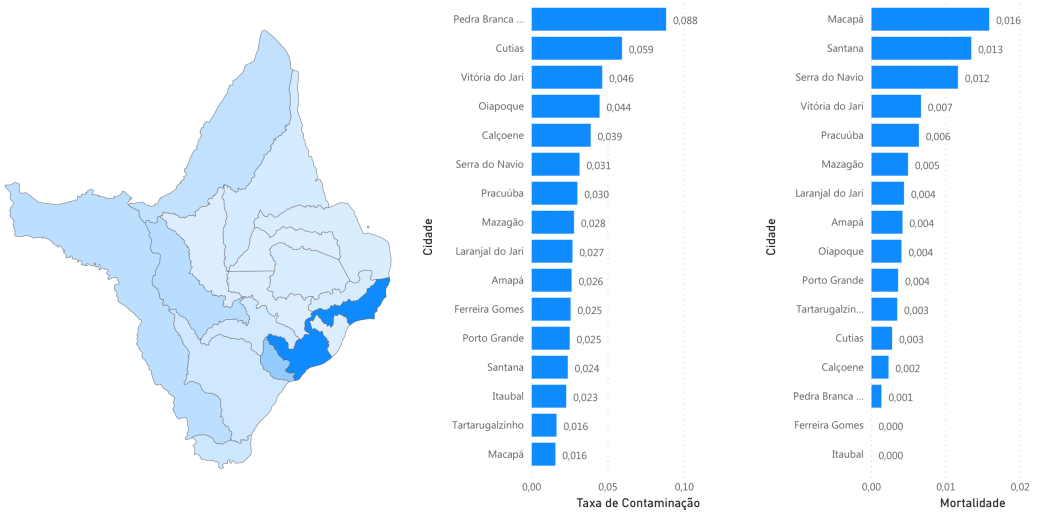
para Junho de 2021 uma maior concentração dos casos na Região Metropolitana de Macapá[3]. Entretanto, para ambos os conjuntos amostrais, a taxa de contaminação é maior em outros municípios, que não fazem parte da Região Metropolitana de Macapá. Por outro lado, a taxa de mortalidade para Junho de 2020 é proporcionalmente maior nos municípios de Macapá, com valor de 0,01, e Santana, também com valor de 0,01, o que não ocorre no conjunto amostral de Junho de 2021, que tem maior taxa de mortalidade para o município de Laranjal do Jari, sendo essa de valor 0.06.

Finalmente foram gerados gráficos comparativos dos dados do estado do Amapá com os outros estados da Região Norte, a saber: Acre, Amazonas, Pará, Rondônia, Roraima e Tocantins. A Figura 7 apresenta os dados referentes ao período de Junho de 2020. Em valores absolutos, o quantitativo de casos notificados é maior no estado do Pará, com valor igual a 67.892, enquanto o Amapá encontra-se em terceiro lugar, com 18.890 notificações (Figura 7-a). Entretanto, ao se calcular a taxa de infecção, que se refere à razão entre o número de notificação e a quantidade de habitantes, o Amapá se enquadra como sendo o estado com maior taxa de contaminação, enquanto o Pará, que possui maior valor absoluto, passa a estar em quinto lugar dentro da região Norte. Apenas Amapá e Roraima ficam acima da média da taxa de mortalidade (linha tracejada no gráfico).

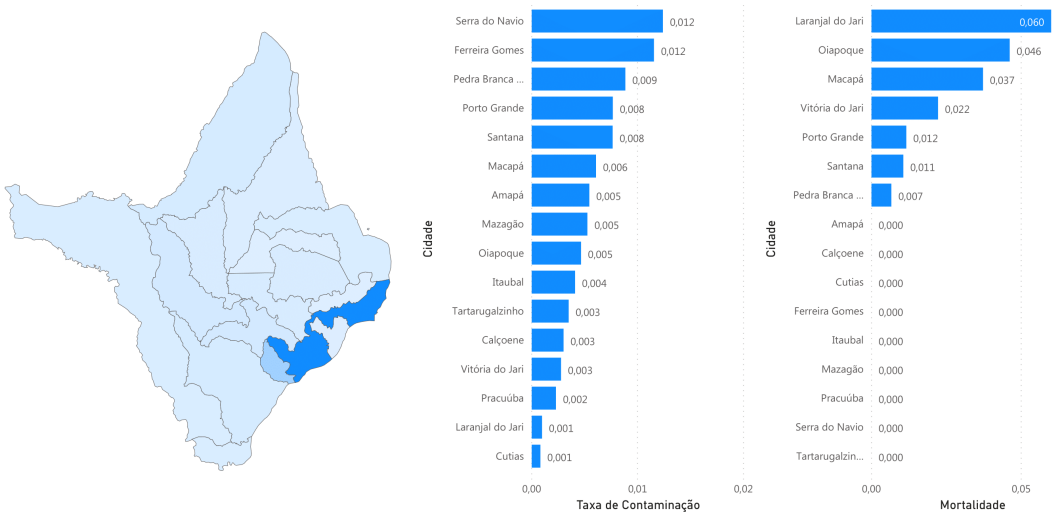
Para a quantidade de óbitos, ainda no período de Junho de 2020, o estado do Pará se apresenta com o maior valor, sendo esta igual a 2.037 registros, enquanto o Amapá encontra-se em sexto lugar, a frente apenas do estado do Tocantins. Por outro lado, em relação à taxa de mortalidade, o estado do Acre fica em primeiro lugar, com valor igual a 0,03. Ficam acima da média os estados do Acre, Pará e Amazonas. O estado do Amapá, embora tenha sido aquele com maior taxa de contaminação, se refere ao estado com menor taxa de mortalidade por Covid-19 para o período considerado.

Para o período de Junho de 2021, de semelhante forma ao observado no período

[3]A Região Metropolitana foi criada pela Lei Complementar Estadual n. 21, de 26 de Fevereiro de 2003, e contempla os municípios de Macapá, Santana e Mazagão. Disponível em: http://www.al.ap.gov.br/ver_texto_consolidado.php?iddocumento=17537. Acesso em 16 Jun, 2021.



(a) Junho de 2020.



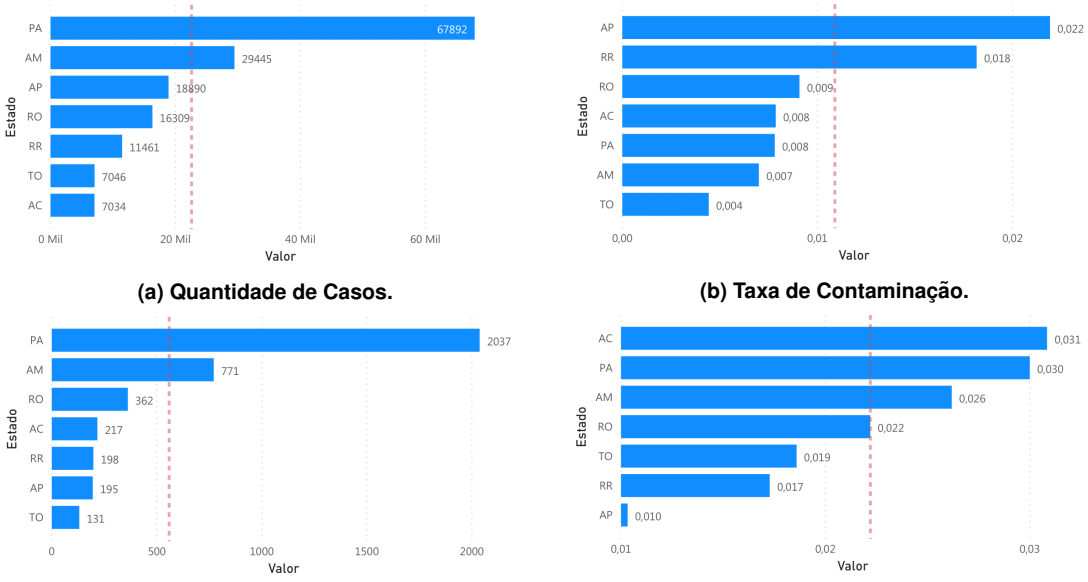
(b) Junho de 2021.

Figura 6. Comparação entre dados de cidades (Amapá): saturação de mapa por número de casos, taxa de contaminação e taxa de mortalidade.

de Junho de 2020, o estado do Pará se enquadra como o de maior número absoluto de notificações de casos de infecção e de registros de óbitos. O estado do Amapá se encontra em sexto lugar, com 5.066 notificações, abaixo da média da região Norte para o período. Para a taxa de contaminação, o maior valor é do estado de Roraima, com 0,01, estando o estado do Amapá abaixo da média da região, em quarto lugar. Ainda neste período temporal, em relação à taxa de mortalidade, o estado do Amapá que em Junho de 2020 apresentou a menor taxa de mortalidade, passou a estar em primeiro lugar em Junho de 2021, com valor igual a 0,02, sucedido pelos estados do Pará e Acre, ambos acima da média regional.

Referências

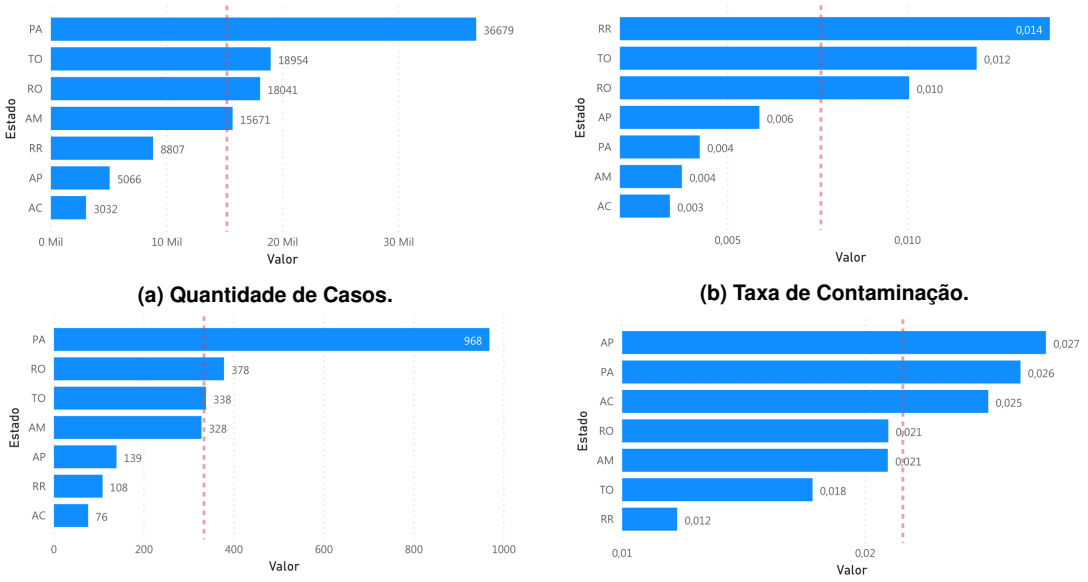
Dean, J. (2014). *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. John Wiley & Sons, New Jersey.



(a) Quantidade de Casos. **(b) Taxa de Contaminação.**

(c) Quantidade de Óbitos. **(d) Taxa de Mortalidade.**

Figura 7. Comparação de valores para região Norte (Junho 2020).



(a) Quantidade de Casos. **(b) Taxa de Contaminação.**

(c) Quantidade de Óbitos. **(d) Taxa de Mortalidade.**

Figura 8. Comparação de valores para região Norte (Junho 2021).

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Pelican, London.