UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

A anatomia dos roubos de veículo no Estado de São Paulo: Uma análise da dinâmica criminal utilizando redes complexas

Vinícius Teixeira de Carvalho Freitas

Trabalho de Conclusão de Curso do Programa de Graduação Bacharelado em Matemática Aplicada e Computação Científica

ICMC USP
SÃO CARLOS

Vinícius Teixeira de Carvalho Freitas

A anatomia dos roubos de veículo no Estado de São Paulo: Uma análise da dinâmica criminal utilizando redes complexas

Trabalho de conclusão de curso apresentado ao Programa de Graduação, do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Bacharel em Matemática Aplicada e Computação Científica.

Área de concentração: Matemática Aplicada e Computação Científica

Orientador: Prof. Dr. Luis Gustavo Nonato

Coorientador: Prof. Dr. Thomas Kauê Dal'Maso Peron

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

T266a

Teixeira de Carvalho Freitas, Vinícius
A anatomia dos roubos de veículos no Estado de São Paulo: Uma análise da dinâmica criminal utilizando redes complexas / Vinícius Teixeira de Carvalho Freitas; orientador Luis Gustavo Nonato; coorientador Thomas Kaue Dal Maso Peron. -- São Carlos, 2023.
100 p.

Trabalho de conclusão de curso (Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Roubo de veículos. 2. Criminalidade. 3. Redes Complexas. 4. Séries Temporais. I. Gustavo Nonato, Luis, orient. II. Kaue Dal Maso Peron, Thomas , coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Vinícius Teixeira de Carvalho Freitas

The Anatomy of Vehicle Theft in the State of São Paulo: An Analysis of Criminal Dynamics Using Complex Networks

Conclusion course paper presented to the Undergraduate Program of the Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, in partial fulfillment of the requirements for the degree of the Bachelor in Applied Mathematics and Scientific Computing.

Concentration area: Applied Mathematics and Scientific Computing

Advisor: Prof. Dr. Luis Gustavo Nonato

Original version

São Carlos

2023

Folha de Aprovação

Autor: Vinícius Teixeira de Carvalho Freitas
Título: A anatomia dos roubos de veículos no Estado de São Paulo: Uma análise da dinâmica criminal utilizando redes complexas
Aprovado em: 18 de dezembro de 2023

Trabalho de conclusão de curso apresentado ao Programa de Graduação, do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Bacharel em Matemática Aplicada e Computação Científica.

Área de concentração: Matemática Aplicada e Computação Científica

Orientador: Prof. Dr. Luis Gustavo Nonato
Coorientador: Prof. Dr. Thomas Kauê Dal'Maso Peron

COMISSÃO JULGADORA:

gov.br Documento assinado digitalmente
JOSE ALBERTO CUMINATO
Data: 25/01/2024 11:11:25-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. José Alberto Cuminato
ICMC - USP
Presidente

gov.br Documento assinado digitalmente
LUIS GUSTAVO NONATO
Data: 25/01/2024 09:01:22-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Luis Gustavo Nonato
ICMC - USP
Membro

gov.br Documento assinado digitalmente
THOMAS KAUE DAL MASO PERON
Data: 26/01/2024 14:38:41-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Thomas Kauê Dal'Maso Peron
ICMC - USP
Membro

São Carlos, 18 de dezembro de 2023

Este trabalho é dedicado àqueles que buscam, com seu próprio talento, superar a humanidade. Às mentes inquietas que pensam, questionam e propõem. Aos que respiram ideias e querem ir além. Eles são a solução.

AGRADECIMENTOS

Gostaria de agradecer, primeiramente, como necessário, ao Todo-Poderoso, por ser imenso na minha vida.

Gostaria também de agradecer aos meus pais, por todo esforço desmedido em prol de meu sucesso e apoio inabalável, aos quais devo tudo que pude conquistar.

Ao contribuinte paulista, figura invisível que torna a Universidade Pública possível a partir de seu árduo trabalho.

Aos amigos João Paulo Clarindo dos Santos, Leandro Narciso Marcelino Rodrigues Gonçalves, e Pedro Balbão Bazon, pela inestimável amizade, apoio constante e momentos compartilhados ao longo desta jornada acadêmica. Se pude evoluir, foi por ter me relacionado com pessoas melhores do que eu.

Aos meus orientadores, Luis Gustavo Nonato e Thomas Kauê Peron, por terem confiado a mim tamanho projeto. E a todos os docentes, discentes e funcionários com quem tive a honra de dividir espaço na Universidade de São Paulo.

"Somos todos professores imperfeitos, mas podemos ser perdoados se tivermos levado a questão um pouco adiante e se tivermos feito o que nos foi possível. Anunciamos o prólogo, e nobremente nos retiramos. Depois de nós, há esperança de que melhores atores virão."

Autor Desconhecido

RESUMO

Freitas, V.T.C **A anatomia dos roubos de veículo no Estado de São Paulo**. 2023. 100p. Monografia (Trabalho de Conclusão de Curso) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Diversas abordagens são empregadas para mitigar as ocorrências de crimes em uma determinada região. Nesse tipo de estudo, grosso modo, é muito importante levar em consideração a distribuição espacial e o padrão temporal das ocorrências. Assim sendo, a presente monografia tem por objetivo utilizar os dados espaço-temporais de ocorrências (roubos de veículos), presentes na transparência do site da Secretaria de Segurança Pública do Estado de São Paulo (SSP-SP), para gerar redes complexas e extrair informações da dinâmica criminal. A metodologia envolve a subdivisão do estado em células, a agregação de eventos criminais nessas células e a criação de séries temporais. A análise de similaridade entre as séries, a partir da medida *Event Synchronization*, resulta na construção de uma rede complexa, em que as células representam vértices e as arestas indicam similaridade entre padrões criminais.

Palavras-chave: Roubo de veículos. Criminalidade. Redes Complexas. Séries Temporais.

ABSTRACT

Freitas, V.T.C **The Anatomy of Vehicle Theft in the State of São Paulo, Brazil**. 2023. 100p. Monograph (Conclusion Course Paper) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Various approaches are employed to mitigate crime occurrences in a specific region. In this type of study, broadly speaking, it is crucial to take into consideration the spatial distribution and temporal pattern of incidents. Therefore, the present thesis aims to use the spatiotemporal data of incidents (vehicle thefts) available on the transparency section of the website of the São Paulo State Public Security Department (SSP-SP) to generate complex networks and extract information regarding criminal dynamics. The methodology involves subdividing the state into cells, aggregating criminal events in these cells, and creating temporal series. The analysis of similarity between the series, using the *Event Synchronization* measure, results in the construction of a complex network where cells represent vertices, and edges indicate similarity between criminal patterns.

Keywords: Vehicle theft. Crime. Complex Networks. Time Series.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

Ingl.	Inglaterra
Finl.	Finlândia
Espan.	Espanha
C.Rica	Costa Rica
Argen.	Argentina
UNICRI	Instituto Inter-regional de Pesquisas das Nações Unidas para o crime e a Justiça
ILANUD	Instituto Latino-Americano das Nações Unidas para Prevenção do Delito e Tratamento do Deliquente
SSP-SP	Secretaria de Segurança Pública do Estado de São Paulo
IML	Instituto Médico Legal
BO	Boletim de Ocorrência
RDO	Registro Digital de Ocorrências
COVID-19	Coronavírus (coronavirus disease 2019)
ES	Event Synchronization

SUMÁRIO

1 INTRODUÇÃO

A criminalidade no Brasil é um desafio multifacetado que tem impactos significativos na sociedade, economia e qualidade de vida dos cidadãos (CHESNAIS, 1999). A complexidade desse problema envolve uma série de fatores, incluindo desigualdades socioeconômicas, falta de acesso à educação de qualidade, deficiências no sistema de justiça e desafios estruturais que demandam uma abordagem holística para sua resolução (CHESNAIS, 1999).

A urgência em enfrentar os problemas relacionados à criminalidade no Brasil é evidente diante das consequências devastadoras que ela acarreta. O país enfrenta altos índices de violência, crimes contra a propriedade, homicídios e tráfico de drogas (Instituto de Pesquisa Econômica Aplicada (Ipea), 2022), o que gera um clima de insegurança generalizado. Além disso, a violência impacta negativamente o desenvolvimento econômico, afetando o turismo, os investimentos e a qualidade de vida da população (MARCHEZINI; SPOLADOR; JORGE, 2020).

A busca por soluções eficazes demanda uma colaboração coesa entre o governo, a sociedade civil e as instituições acadêmicas (DIAS, 2019). A implementação de políticas públicas efetivas, informadas por pesquisas aprofundadas e estudos na área da criminalidade, se torna imperativa. Investir em pesquisas que compreendam as causas subjacentes do fenômeno criminal é essencial para o desenvolvimento de estratégias (GUEDES, 2007). Essa abordagem baseada em evidências não apenas resgataria a sensação de segurança na população, mas também contribuiria para o progresso e a prosperidade do país, estabelecendo um ambiente propício para o avanço sustentável.

1.1 Motivação

Crimes envolvendo veículos impactam diretamente na mobilidade e na sensação de segurança da população (DIXON; FARRELL, 2020), tornando-se um objeto de estudo relevante para o desenvolvimento de estratégias eficazes de combate à problemática e para a promoção da segurança pública, em que a tecnologia, como a análise de dados, desempenha um papel crucial para o sucesso das investigações (WALSH; TAYLOR, 2007). Roubos e furtos de veículos têm semelhanças em suas naturezas e motivações; porém, o roubo é, muitas vezes, mais sintomático devido à presença de ameaças e coação[1], frequentemente envolvendo o uso de armas. Em contrapartida, o furto ocorre de maneira não coerciva, diferenciando-se pela ausência de ameaças diretas à integridade física [2], o que influencia substancialmente a dinâmica e o impacto psicológico para as vítimas.

Além disso, o roubo de veículos impõe desafios específicos às forças de segurança pública, exigindo consideráveis recursos humanos, tecnológicos e de infraestrutura para a efetiva recuperação dos automóveis subtraídos (Portal do Governo do Estado de São Paulo, 2012). Há também uma conexão intrínseca entre o roubo de veículos e redes de crime organizado (WALLACE,

[1] O roubo é descrito no artigo 157 do Código Penal, sendo caracterizado pela subtração de bem material mediante grave ameaça ou violência.

[2] O furto é descrito no artigo 155 do Código Penal, é caracterizado pela subtração de bem material alheio (destacando-se a ausência de ameaça grave ou violência).

2004), que utilizam os veículos roubados para diversas atividades, tais como o transporte de drogas, contrabando de mercadorias ilícitas e assaltos à mão armada. Combatê-lo implica também desmantelar parte das operações dessas organizações criminosas, contribuindo, assim, para a redução da criminalidade.

A expectativa de que as vítimas notifiquem essas ocorrências de maneira mais consistente, diminuindo os índices de subnotificação (c.f. Tabela 1), especialmente devido à necessidade de registro para fins de seguro e recuperação de veículos roubados, cria uma base de dados robusta e confiável a partir de boletins de ocorrência, diferentemente de outros tipos de crimes, como os contra a vida, que enfrentam desafios na notificação e perícia (COSTA, 2021).

Tabela 1 – Taxa de Notificação - Cidades com mais de 100 mil habitantes de países selecionados, 1992

Tipo de Crime	**Ingl.**	**Finl.**	**Espan.**	**Itália**	**C.Rica**	**Brasil**	**Argen.**
Roubo de carro	93,9	100,0	80,9	94,9	73,7	91,9	90,3
Furto de dentro do carro	74,3	55,0	29,2	40,1	22,1	18,3	53,8
Vandalismo no carro	35,5	36,1	18,4	14,9	18,2	0,9	18,8
Roubo de moto	93,5	85,7	85,4	76,4	91,7	65,0	79,5
Roubo de bicicleta	74,6	54,6	40,9	27,5	35,7	7,1	41,4
Arrombamento	94,6	75,0	70,8	65,5	50,8	38,4	68,9
Tentativa de arrombamento	55,2	22,2	22,5	20,9	22,5	19,3	40,9
Assalto	52,1	28,6	32,1	37,5	27,6	19,1	42,0
Ofensas sexuais	16,4	11,2	3,6	4,3	9,3	9,8	43,0
Agressão/ameaça	41,7	24,4	24,4	25,4	29,9	11,5	34,4

Fonte: UNICRI / ILANUD

1.2 Objetivos

O estudo proposto visa extrair informações dos dados de boletins de ocorrência relacionados a crimes envolvendo veículos, em particular, o roubo de veículos, no Estado de São Paulo, a partir de redes complexas. Essas informações estão acessíveis na transparência do site da Secretaria de Segurança Pública[3] (SSP-SP). A metodologia envolve a subdivisão do estado em células, a agregação de eventos criminais nessas células e a criação de suas séries temporais. A análise de similaridade entre as séries, a partir da medida Event Synchronization, resulta na construção de uma rede complexa, em que as células representam vértices e as arestas indicam similaridade entre padrões criminais.

A ideia de modelar a dinâmica criminal dessa forma surge porque a complexidade do sistema de crimes em uma região é influenciada por fatores macroeconômicos (BOTHOS; THOMOPOULOS, 2016), governamentais, legislativos, judiciais, sociais e culturais (HAINES, 1999; WAWRZYNIAK *et al.*, 2018). A interconexão desses elementos, agindo em diferentes escalas espaciais e temporais, sugere a abordagem de redes complexas como uma ferramenta eficaz para modelar e compreender as interações que contribuem para o fenômeno.

[3] https://www.ssp.sp.gov.br

Um número significativo de pesquisas (AMISANO, 2018; JUNIOR, 2021; SANTOS, 2019; SAMPAIO, 2023) dedicou-se à análise dos fatores que permeiam a ação criminal, fazendo com que componentes desse sistema tenham sido identificados e descritos na literatura. Essas pesquisas estão ganhando crescente viabilidade (PAUW, 2011), em parte devido ao acesso facilitado a dados criminais, uma conquista impulsionada pela era da informação. Contudo, ainda há diversos desafios relacionados à previsão da mudança em frequência e extensão de padrões para ocorrência de delitos, fator crucial na mitigação das consequências da dinâmica criminal (WALSH; TAYLOR, 2007). O presente trabalho se propõe a estabelecer conexões entre diferentes regiões geoespaciais, representadas por células, ao analisar seus padrões criminais e relacionar suas ocorrências. O objetivo é extrair informações sobre a estrutura dos roubos de veículos no Estado de São Paulo e estender métodos para previsão em mudanças nos padrões para a ocorrência de delitos.

A escolha por utilizar dados da SSP-SP como representação dos registros do Estado de São Paulo é respaldada por diversos motivos. Os dados fornecidos pela SSP-SP seguem padrões de coleta e armazenamento, o que garante a qualidade e confiabilidade das informações. Além disso, a boa acessibilidade aos dados simplifica o processo de obtenção. A documentação adequada sobre seu manuseio permite que pesquisas científicas sejam realizadas a partir deles. Ressalta-se que, embora o estudo tenha se concentrado no Estado de São Paulo, a metodologia desenvolvida pode ser aplicada a dados de outros estados da União.

O texto foi subdividido entre as informações pertinentes acerca dos dados utilizados, a teoria necessária para compreensão do trabalho, a metodologia empregada para a obtenção dos resultados, e os resultados e discussões subsequentes. No Capítulo 2 há uma descrição das informações que estão sendo utilizadas. Lá consta onde é possível encontrá-las, o sistema e processo pelo qual elas são aferidas e ratificadas, os critérios utilizados para sua correta discriminação, os detalhes de cunho jurídico relacionados ao crime cometido, uma metodologia para sua interpretação e os procedimentos prescritos para a análise, além das limitações de acesso provenientes da lei geral de dados. Também constam os meios para sua obtenção e estruturação.

O Capítulo 3 traz uma fundamentação teórica sobre os conceitos mais relevantes para o desenvolvimento do trabalho, indicando bibliografias que possam complementar o conteúdo. Constam breves conceituações de séries temporais, séries de eventos, medidas de similaridade, *Event Synchronization* e redes complexas.

No Capítulo 4, apresenta-se de forma detalhada a metodologia delineada para a seleção dos dados do trabalho, a determinação das células e a criação do *grid* que subdivide o estado, bem como os procedimentos para a criação e filtragem das séries temporais que representam a atividade criminal de cada célula. Além disso, são apresentadas as escolhas para determinar a similaridade entre as séries, o que norteia, posteriormente, a criação das redes complexas que foram estudadas.

No Capítulo 5, apresentam-se os resultados e conclusões em tabelas e gráficos que ilustram as descobertas obtidas. Além disso, são exploradas as possíveis causas por trás desses resultados, proporcionando interpretações da teoria discutida. Destacar-se-á uma análise comparativa distintos tipos de redes desenvolvidas, oferecendo uma compreensão aprofundada de suas características individuais e impactos.

A versão digital da monografia, que possui interessantes recursos para auxiliar a leitura, está disponível na Biblioteca Digital de Trabalhos Acadêmicos da USP[4]. Os programas para gerar as figuras, para extração dos dados, para o tratamento e seleção dos dados, e para a aplicação da metodologia e obtenção dos resultados do trabalho estão disponíveis mediante solicitação ao autor.

[4] https://bdta.abcd.usp.br

2 OS DADOS

A SSP-SP declara possuir o maior portal de informações sobre segurança pública do país, que teve início em 9 de maio de 2016. Nele, estão presentes: **Taxa de Homicídio**, que oferece acesso à série histórica das taxas anuais de homicídio por regiões do estado; **Registro de Óbitos – IML**, que disponibiliza informações básicas sobre todas as entradas de óbitos no IML; **Boletins de Ocorrência**, a partir do sistema de Registro Digital de Ocorrências (RDO).

Para a presente monografia, são de interesse apenas os dados contidos no sistema RDO, que concedem acesso a boletins de ocorrência emitidos dentro do Estado de São Paulo desde o ano de 2003. As seguintes naturezas consumadas estão disponíveis: homicídio doloso, roubo seguido de morte (Latrocínio), lesão corporal seguida de morte, morte decorrente de intervenção policial, morte suspeita, roubo de veículo, furto de veículo, roubo de celular e furto de celular. No presente capítulo, a apresentação foi direcionada para a natureza consumada "roubo de veículo", foco da análise. Mais informações pertinentes às demais naturezas podem ser encontradas na apresentação do site da Secretaria de Segurança Pública do Estado de São Paulo[1].

2.1 Obtenção dos dados

A aquisição dos dados utilizados para o trabalho foi realizada através do site da transparência da SSP-SP[2] (c.f. Figura 1). O portal disponibiliza acesso direto a todas as tabelas referentes às naturezas consumadas mencionadas no início deste capítulo, além de informações sobre como interpretá-las.

Não há um padrão geral para a extração dos dados de cada tabela. Para baixar todo o conteúdo presente das naturezas consumadas: *homicídio doloso*, *feminicídio*, *latrocínio*, *lesão corporal seguida de morte* e *morte decorrente de intervenção policial*, basta selecionar os respectivos botões. Além de todo período disponível vir em um só arquivo, este está em um formato ".xlsx", mais moderno e eficiente.

Já para *registro de óbitos - IML*, *morte suspeita*, *furto de veículo*, *roubo de veículo*, *furto de celular*, *roubo de celular* e *SP dados criminais*, é necessário escolher o ano e mês indicados (c.f Figura 2), o que implica baixar mês a mês, ano a ano, cada planilha que compreenda o período que se pretende estudar, processo agravado por um alto tempo requerido pelo site para encontrar os documentos em seu banco de dados. O formato dos arquivos é ".xls", obsoleto, com várias planilhas corrompidas ou com erros sendo fornecidas pelo site. O processo de extração empreendido pelo trabalho para os dados pertencentes a "roubo de veículo" foi descrito em detalhes no Capítulo 4.

É importante destacar que as tabelas apresentam valores ausentes ou inconsistentes, uma condição que pode ser atribuída a uma variedade de fatores, incluindo possíveis falhas durante o processo de coleta, erros na entrada de dados e características inerentes à natureza do próprio

[1] https://www.ssp.sp.gov.br/transparenciassp/Apresentacao.aspx
[2] http://www.ssp.sp.gov.br/transparenciassp/Consulta2022.aspx

Figura 1 – O portal da transparência da SSP-SP

Registro feito em 10 set. 2023.

conjunto de dados (MIOT, 2019). Valores ausentes, grosso modo, impossibilitam a aplicação de metodologias e podem adicionar viés ao estudo (HYUN, 2013). Assim, torna-se essencial ponderar e adotar estratégias para mitigar a ausência de valores ou inconsistências.

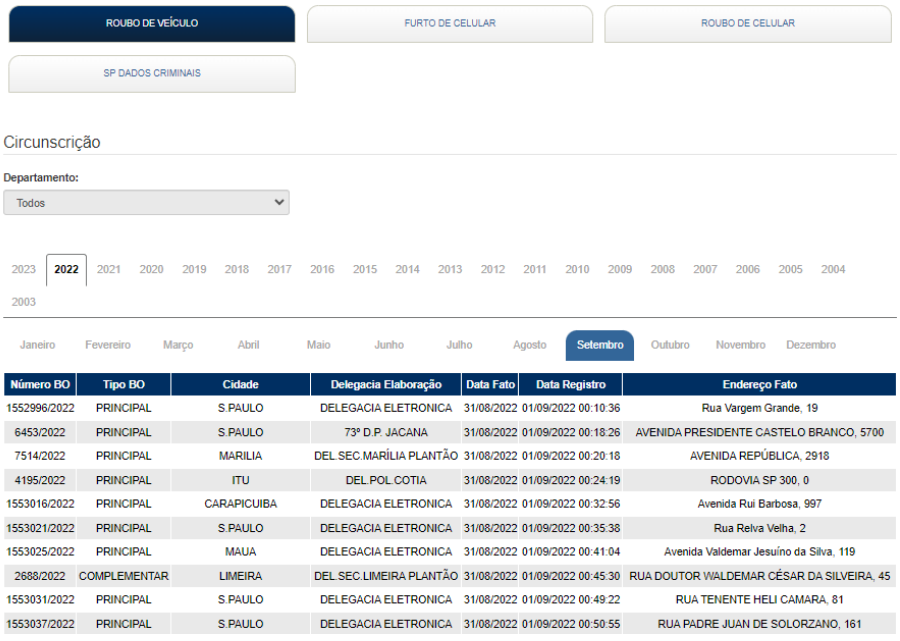
Um estudo recente (FREITAS; CLARINDO; AGUIAR, 2023) abordou integralmente o processo de extração e estruturação, resultando na criação de um banco de dados que abrange todas as planilhas disponíveis na transparência da SSP-SP. No entanto, esse banco não foi utilizado no presente trabalho, uma vez que seu desenvolvimento e publicação ocorreram posteriormente à conclusão das etapas de extração, estruturação e seleção realizadas no presente estudo.

2.2 Introdução aos dados da SSP-SP

A SSP-SP oferece um documento destinado à leitura e interpretação dos dados, acessível por meio do botão ”Exportar Metodologia” destacado na Figura 2b. Abaixo, são listados os principais pontos contidos neste documento.

- Os dados constantes foram extraídos do sistema de Registro Digital de Ocorrências (RDO) que é a ferramenta de registro dos boletins de ocorrência nas delegacias de polícia. Para todas as tabelas, cada linha representará um boletim de ocorrência emitido dentro da jurisdição do Estado de São Paulo.

Figura 2 – Opções para seleção e download



Número BO	Tipo BO	Cidade	Delegacia Elaboração	Data Fato	Data Registro	Endereço Fato
1552996/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 00:10:36	Rua Vargem Grande, 19
6453/2022	PRINCIPAL	S.PAULO	73º D.P. JACANA	31/08/2022	01/09/2022 00:18:26	AVENIDA PRESIDENTE CASTELO BRANCO, 5700
7514/2022	PRINCIPAL	MARILIA	DEL SEC.MARÍLIA PLANTÃO	31/08/2022	01/09/2022 00:20:18	AVENIDA REPÚBLICA, 2918
4195/2022	PRINCIPAL	ITU	DEL.POL.COTIA	31/08/2022	01/09/2022 00:24:19	RODOVIA SP 300, 0
1553016/2022	PRINCIPAL	CARAPICUIBA	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 00:32:56	Avenida Rui Barbosa, 997
1553021/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 00:35:38	Rua Relva Velha, 2
1553025/2022	PRINCIPAL	MAUA	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 00:41:04	Avenida Valdemar Jesuíno da Silva, 119
2688/2022	COMPLEMENTAR	LIMEIRA	DEL.SEC.LIMEIRA PLANTÃO	31/08/2022	01/09/2022 00:45:30	RUA DOUTOR WALDEMAR CÉSAR DA SILVEIRA, 45
1553031/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 00:49:22	RUA TENENTE HELI CAMARA, 81
1553037/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 00:50:55	RUA PADRE JUAN DE SOLORZANO, 161

(a) Seleção de ano e mês

1553037/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 00:50:55	RUA PADRE JUAN DE SOLORZANO, 161
1198/2022	PRINCIPAL	TABOAO DA SERRA	02º D.P. TABOÃO DA SERRA	31/08/2022	01/09/2022 01:06:34	
4629/2022	COMPLEMENTAR	S.PAULO	24º D.P. PONTE RASA	25/08/2022	01/09/2022 01:13:19	Avenida Capitão Anselmo Barcelos, 558
4168/2022	COMPLEMENTAR	COTIA	DEL.POL.COTIA	29/08/2022	01/09/2022 01:15:29	RUA MÁISA, 1
6455/2022	COMPLEMENTAR	S.PAULO	73º D.P. JACANA	31/08/2022	01/09/2022 01:27:02	AVENIDA EDUCADOR PAULO FREIRE, 815
2087/2022	COMPLEMENTAR	PRAIA GRANDE	01º D.P. SÃO VICENTE	31/08/2022	01/09/2022 01:34:45	Rua Tavares Bastos, 462
1553184/2022	PRINCIPAL	S.PAULO	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 01:36:27	Avenida Raimundo Pereira de Magalhães, 16250
1553183/2022	PRINCIPAL	GUARULHOS	DELEGACIA ELETRONICA	31/08/2022	01/09/2022 01:36:27	RUA ENVIRA, 174
3295/2022	PRINCIPAL	S.BERNARDO DO CAMPO	03º D.P. DIADEMA	31/08/2022	01/09/2022 02:02:26	Avenida Horácio Barione, 296
3424/2022	PRINCIPAL	S.PAULO	37º D.P. CAMPO LIMPO	31/08/2022	01/09/2022 02:22:13	RUA JOSE NOGUEIRA, 100
2927/2022	PRINCIPAL	S.ANDRE	02º D.P. SANTO ANDRÉ	31/08/2022	01/09/2022 03:15:42	Avenida Presidente Costa e Silva, 341

(b) Botão ”Exportar” realiza download

- O sistema RDO teve sua implantação concretizada de modo gradual nas diversas unidades policiais do Estado, alcançando todos os municípios apenas a partir do ano de 2010.

- Os boletins são apresentados conforme foram registrados pelas unidades policiais, no sistema RDO. Isso significa que eventuais erros, como valores ausentes ou inconsistentes, vêm diretamente do momento de confecção dos boletins.

- O número total de boletins de ocorrência registrados sob uma natureza consumada não representa a estatística criminal do estado ou de determinada área ou região. Isso se deve, entre outros fatores, à subnotificação.

- A inclusão ou alteração de um campo e respectivos períodos de implementação podem influenciar diretamente nos critérios de pesquisa executados, não havendo na base fornecida o tratamento metodológico necessário para qualificá-las como dados estatísticos oficiais.

- Cada linha constante na tabela registra os dados de uma pessoa, natureza ou objeto relacionado no boletim. Assim, um boletim que possua a identificação de mais de uma pessoa, natureza ou objeto (a depender da pesquisa solicitada) terá os dados da ocorrência multiplicados pelos indexadores solicitados, ou seja, várias linhas podem se referir ao mesmo boletim.
- Para conclusões quanto às quantidades nominais de ocorrências, é necessária a exclusão das duplicidades por meio dos campos: NOME_DELEGACIA, ANO_BO, NUM_BO.
- Boletins que envolvem múltiplas naturezas, disponíveis no portal, serão apresentados em ambas as categorizações. Isto é, um mesmo incidente pode estar presente em tabelas de diferentes naturezas se ele contempla mais de um tipo de crime.

O documento ressalta, adicionalmente, a presença de restrições legais que afetam as informações veiculadas. Essas limitações estão associadas a questões jurídicas e regulatórias, impondo condições específicas sobre a divulgação e acessibilidade de determinadas informações.

- Não serão fornecidos históricos de quaisquer naturezas do Título VI - Dos Crimes Contra os Costumes/Dignidade sexual e do Título I - Dos Crimes Contra a Pessoa Capítulo V - Dos Crimes Contra a Honra do Código Penal.
- São protegidas as ocorrências que tenham associadas quaisquer naturezas do Título VI, do Código Penal (Crimes contra os Costumes/Dignidade Sexual), suprimindo-se o nome da vítima.
- Os campos que podem levar à identificação da pessoa são protegidos de acordo com o art. 31 da Lei de Acesso a Informação.
- Não são disponibilizados endereços quando o tipo de local tiver sido registrado como residência ou congênere.
- Em virtude das características das informações contidas nos históricos, que tem por finalidade principal a coleta de subsídios para o início das investigações pelas autoridades policiais, além de conterem dados pessoais, para acesso aos históricos deverá ser atendido o previsto no art. 31 da Lei nº 12.527, de 18 de novembro de 2011, em especial o §3º:

> Art. 31. O tratamento das informações pessoais deve ser feito de forma transparente e com respeito à intimidade, vida privada, honra e imagem das pessoas, bem como às liberdades e garantias individuais.
>
> §1. As informações pessoais, a que se refere este artigo, relativas à intimidade, vida privada, honra e imagem:
>
> II - poderão ter autorizada sua divulgação ou acesso por terceiros diante de previsão legal ou consentimento expresso da pessoa a que elas se referirem.
>
> §2. Aquele que obtiver acesso às informações de que trata este artigo será responsabilizado por seu uso indevido.

§3. O consentimento referido no inciso II do §1 não será exigido quando as informações forem necessárias:

I - à prevenção e diagnóstico médico, quando a pessoa estiver física ou legalmente incapaz, e para utilização única e exclusivamente para o tratamento médico;

II - à realização de estatísticas e pesquisas científicas de evidente interesse público ou geral, previstos em lei, sendo vedada a identificação da pessoa a que as informações se referirem;

III - ao cumprimento de ordem judicial;

IV - à defesa de direitos humanos; ou

V - à proteção do interesse público e geral preponderante.

2.3 Dados relativos a veículos

Para esclarecer as operações deste trabalho, faz-se relevante destacar alguns detalhes das planilhas referentes a roubos de veículos. Seus dados abrangem o período de 2003 a 2022. No entanto, o sistema RDO não estava completamente implementado em todo o estado até 2010, como mencionado anteriormente. Portanto, os registros fora das regiões metropolitanas possivelmente não são representativos, pois não abrangem todas as delegacias do estado.

No total, são 240 tabelas, 12 para cada ano, cada uma representando um mês. A quantidade total de instâncias, juntando todas as planilhas, é de 2.568.107. Removendo-se as entradas duplicadas, um total de 1.376.519 roubos de veículo distintos foram registrados no período. Para o ano de 2022, em específico, constam 124.512 instâncias, com 52.554 ocorrências sendo distintas entre si.

As tabelas brutas extraídas do site da SSP-SP possuem 54 colunas representando informações a serem registradas durante a confecção do BO. Um fator inesperado é a presença de informações que são pertinentes a crimes envolvendo celulares. A SSP-SP não faz qualquer menção a isso em seu documento de interpretação, mas possivelmente foi usado o mesmo *template* para ambos os contextos por envolverem naturezas afins (furto ou roubo).

Um quadro com todas as colunas das planilhas, suas respectivas descrições e tipo ideal está no Apêndice A. O Quadro 1 destaca as colunas mais relevantes para o presente trabalho, usadas no Capítulo 4 para obter os resultados. Os demais campos não foram usados diretamente.

Os três primeiros campos são necessários para a exclusão de entradas duplicadas, conforme indicado na Seção 2.2. As colunas 'DATAOCORRENCIA' e 'HORAOCORRENCIA' fornecem informações temporais das ocorrências, fator crucial para identificar o padrão criminal. Os campos subsequentes, 'BAIRRO' e 'CIDADE', apresentam informações espaciais categóricas sobre unidades administrativas que localizam a ocorrência. Por fim, 'LATITUDE' e 'LONGITUDE' fornecem informações espaciais e geográficas, que permitem localizar com precisão o ponto exato em que o delito se sucedeu.

Quadro 1 – Principais campos contidos nas tabelas de roubo de veículo

Nome da Coluna	**Descrição**	**Tipo de Dado (ideal)**
ANO_BO	Ano de identificação do BO	Inteiro
NUM_BO	Número de identificação do BO	Inteiro
DELEGACIA_NOME	Nome da delegacia de registro	Texto
DATAOCORRENCIA	Data da ocorrência	Data
HORAOCORRENCIA	Hora da ocorrência	Hora
BAIRRO	Bairro da ocorrência	Texto
CIDADE	Cidade da ocorrência	Texto
LATITUDE	Latitude da ocorrência	Ponto flutuante
LONGITUDE	Longitude da ocorrência	Ponto flutuante

Fonte: Elaborado pelo autor.

3 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica tem como objetivo contextualizar os principais temas para a compreensão deste trabalho: séries temporais, com ênfase em séries de eventos; medidas de similaridade entre séries temporais, com foco em *Event Synchronization*; e Redes Complexas.

3.1 Séries temporais

A análise de dados experimentais observados em momentos distintos no tempo traz consigo novos e singulares desafios na modelagem estatística e inferência "tradicionais". A correlação evidente introduzida pela amostragem de pontos de tempo adjacentes pode restringir significativamente a aplicabilidade de muitos métodos estatísticos convencionais, que tradicionalmente dependem da suposição de que essas observações adjacentes são independentes e identicamente distribuídas. A abordagem sistemática para lidar com as questões matemáticas e estatísticas levantadas por essas correlações temporais é comumente conhecida como análise de séries temporais (WEI, 1994)

O impacto da análise de séries temporais em aplicações científicas pode ser parcialmente documentado ao produzir uma lista resumida dos diversos campos nos quais problemas importantes de séries temporais podem surgir (WEI, 1994). Por exemplo, muitas séries temporais conhecidas ocorrem no campo da economia, onde há continua exposição de cotações diárias do mercado de ações ou dados mensais de desemprego. Cientistas sociais acompanham séries sobre a população de um determinado lugar, como datas de nascimento ou matrículas escolares. Um epidemiologista pode estar interessado no número de casos de COVID-19 observados ao longo de algum período de tempo. O governo pode estar interessado em várias séries temporais de padrões criminais que poderiam ser estudadas para prevenir futuros delitos.

3.1.1 Noções gerais

Uma série temporal pode ser intuitivamente definida como qualquer conjunto de observações ordenadas no tempo (c.f Figura 3). São exemplos de séries temporais:

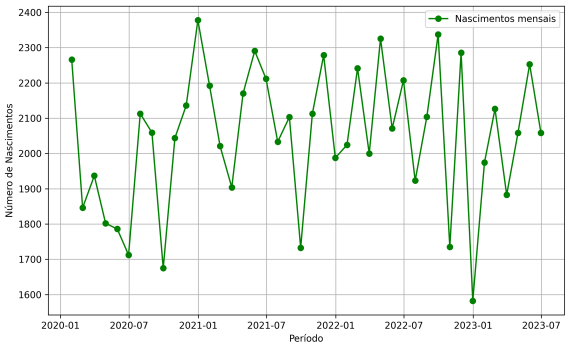
- Os índices diários da Bolsa de Valores de São Paulo (como o Ibovespa) são séries temporais que rastreiam o desempenho das ações e ativos financeiros ao longo do tempo. Esses índices são calculados com base nas flutuações dos preços das ações negociadas na bolsa e são atualizados diariamente. Eles são usados para acompanhar a evolução dos mercados financeiros e tomar decisões de investimento.

- A quantidade mensal de nascimentos em uma determinada região é uma série temporal cujas observações são o valor nominal da natalidade e cada observação está condicionada a um determinado mês de registro.

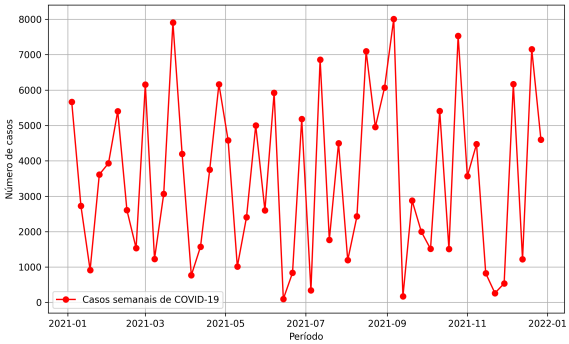
- A quantidade semanal de novos casos positivos para COVID-19 é uma série temporal cujas observações são o total de casos nos sete dias e elas estão dispostas conforme a sequência das semanas.
- O total de ocorrências diárias de roubos de veículos em uma área específica é uma série temporal cujas observações são ocorrências deste tipo de crime num determinado dia, e elas estão dispostas conforme a sequência dos dias.
- O registro de um evento sísmico é uma série temporal cujas observações são os movimentos de vibração provocados e ele está disposto continuamente durante todo o registro do evento.

Figura 3 – Exemplos gráficos de séries temporais

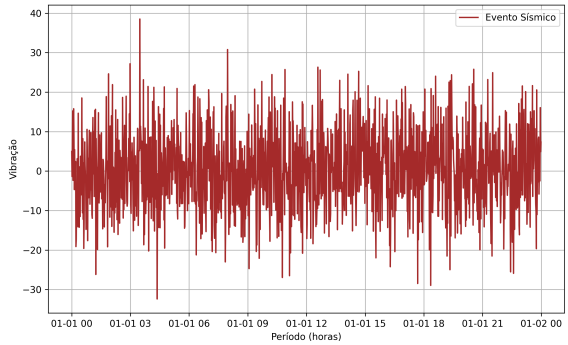
(a) Exemplo de série financeira



(b) Exemplo de série de natalidade



(c) Exemplo de série epidemiológica



(d) Exemplo de série sísmica

Séries ilustrativas geradas pelo autor

Nos exemplos listados, as quatro primeiras séries possuem seu intervalo de tempo discreto; séries temporais de tempo discreto referem-se a conjuntos de dados onde as observações são registradas em intervalos de tempo definidos e distintos, como dias, semanas ou meses. O tamanho do intervalo de tempo é dito o *lag* ou defasagem da série e a quantidade de intervalos é chamada de *timesteps*, passos de tempo ou resolução. A última série possui intervalo de tempo contínuo, pois a atividade sísmica não acontece em intervalos, mas continuamente no tempo. Muitas vezes, uma série temporal discreta é obtida através da amostragem de uma série temporal contínua em intervalos de tempos iguais (MORETTIN, 2006).

Além disso, os valores assumidos pela séries, ou observações, podem ser discretos (valores enumeráveis) ou contínuos (números reais). Por estarem limitados aos centavos (centésimos), pode-se classificar os índices da bolsa como discretos, entretanto, o tratamento destes muitas vezes será como dados contínuos. As quantidades descritas nos exemplos anteriores representam contagens e, portanto, os valores são discretos. Contudo, o último exemplo representa uma série de valores essencialmente contínuos.

Além dessa categorização, é útil compreender as nuances entre séries temporais univariadas e multivariadas, unidimensionais e multidimensionais. Séries temporais univariadas registram as flutuações de apenas uma quantidade ao longo do tempo, como as séries mencionadas anteriormente. Por outro lado, as series temporais multivariadas registram simultaneamente várias quantidades (MORETTIN, 2006). Por exemplo, uma série que represente altura, temperatura e pressão de um gás é considerada multivariada, pois contempla múltiplas variáveis ao mesmo tempo.

No contexto da dimensionalidade, as séries unidimensionais são indexadas em um único fator, geralmente o tempo. Todas as séries previamente discutidas são unidimensionais. Por outro lado, séries temporais multidimensionais possuem mais de um fator indexador, podendo incluir, por exemplo, tempo, latitude e longitude. Uma série que realize a medição da temperatura e pressão geoespacial dos oceanos é um exemplo de série multivariada e multidimensional. [1]

Lidar com as diversas naturezas das séries temporais é um desafio em si, uma vez que os conceitos gerais, tais como padrões, ciclos, autocorrelação, entre outros, variam de acordo com a forma como a série registra e assume seus valores. Isso, por sua vez, influencia diretamente os métodos utilizados para estudar, modelar e prever essas séries. Os procedimentos descritos a seguir idealizam séries com intervalo de tempo discreto, ou ao menos que possam ser transformadas para tal, sem prejuízo da análise. Mais detalhes em Morettin (2006).

3.1.2 Definição formal

Ao discutir séries temporais, é essencial conceituar variável aleatória e processo estocástico. Uma variável aleatória é uma formalização matemática de uma quantidade ou objeto sujeito a eventos aleatórios, que consiste em uma função dos resultados possíveis em um espaço amostral para um espaço mensurável, muitas vezes associado aos números reais. Sua definição formal é dada por:

Seja $(\Omega, \mathcal{F}, \mathcal{P})$ um espaço de probabilidade para o espaço amostral Ω, a σ-álgebra $\mathcal{F}$ e a função de probabilidade $\mathcal{P}$. Denomina-se variável aleatória qualquer função $X : \Omega \to \mathbb{R}$ tal que

$$X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{F},$$

para todo intervalo $I \subset \mathbb{R}$. Pode-se entender uma série temporal como uma sequência de observações ordenadas ao longo do tempo, em que cada observação é uma realização de uma variável aleatória em um ponto específico desse intervalo temporal. Em outras palavras, para cada instante de tempo na série, é obtida uma variável aleatória associada, representando a aleatoriedade inerente aos dados observados: considerando a série temporal $X(t)$, onde t pertence

[1] Neste estudo, o interesse reside em séries temporais univariadas e unidimensionais.

ao conjunto T de instantes de tempo, cada $X(t)$ é uma variável aleatória. Para uma compreensão mais aprofundada, recomenda-se a consulta ao capítulo 2 do livro de Magalhães (2006).

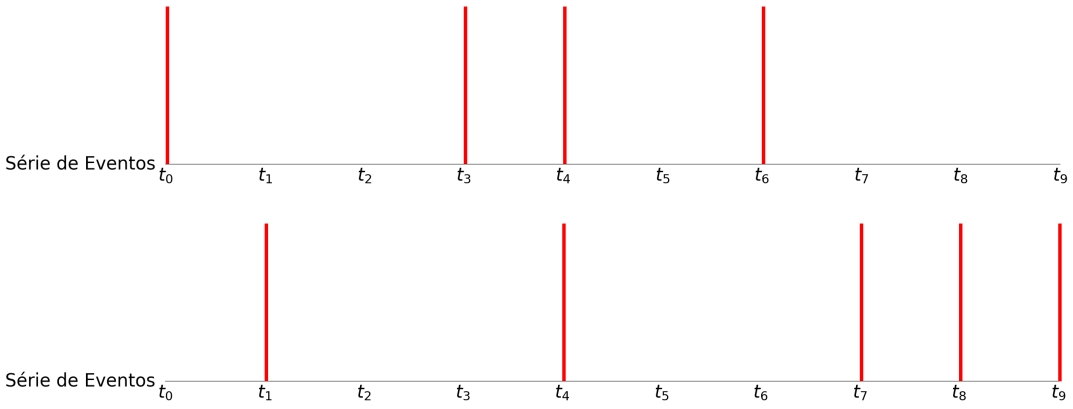
Agora, considerando um conjunto arbitrário T, um processo estocástico é uma família $Z = \{Z(t)\,,\ t \in T\}$ tal que, para cada $t \in T$, $Z(t)$ é uma variável aleatória. Em outras palavras, um processo estocástico é uma família de variáveis aleatórias definidas no mesmo espaço de probabilidades $(\Omega, \mathcal{A}, \mathcal{P})$. Dado que, para $t \in T$, $Z(t)$ é uma variável aleatória definida sobre Ω, na realidade $Z(t)$ é uma função de dois argumentos, $Z(t, \omega)$, $t \in T$, $\omega \in \Omega$. Mais detalhes sobre processos estocásticos e séries temporais podem ser encontrados em Shumway e Stoffer (2006).

Assim sendo, pode-se encarar toda a série temporal como um processo estocástico. Precisamente, uma série temporal representa apenas uma entre as diversas realizações possíveis de um processo estocástico (MORETTIN, 2006). Ao considerar um processo estocástico como uma coleção de variáveis aleatórias indexadas ao longo do tempo, cada série temporal específica corresponde a uma trajetória particular desse processo no espaço amostral, dado que cada ponto na série temporal é uma observação singular de uma variável aleatória associada a um instante específico no tempo. A análise de séries temporais no contexto de variáveis aleatórias e processos estocásticos permite identificar tendências e realizar previsões, levando em consideração a natureza probabilística dos eventos ao longo do tempo.

3.1.3 Séries temporais de eventos

Séries temporais de eventos, ou séries de eventos, são séries temporais binárias, isto é, cujo valor da variável aleatória assume apenas 0 ou 1, e em que valores não nulos representam um evento, como ilustra a Figura 4. Sua representação é muito simples e pode ser visualizada a partir dos seus eventos.

Figura 4 – Exemplos de séries de eventos com nove passos de tempo $(t_1, ..., t_9)$



Séries ilustrativas geradas pelo autor

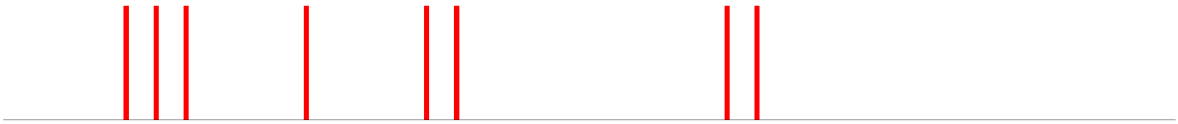
Existem muitas opções possíveis para a definição de evento. Por exemplo, os eventos podem ser momentos no tempo em que o valor da série temporal está acima de um limiar global, ou em que esse valor está acima de um percentil específico da distribuição dos valores,

ou momentos no tempo em que os valores da série mudam drasticamente. Como essa escolha depende da pergunta de pesquisa específica e da aplicação, a definição de eventos pode variar bastante (BOERS, 2015).

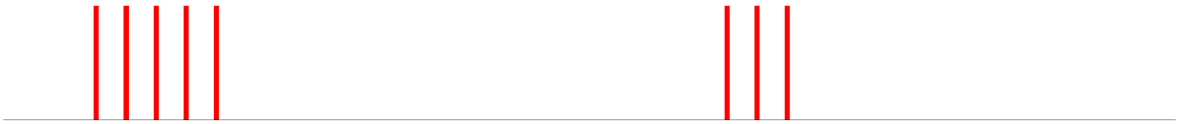
Uma série de eventos de alta resolução é caracterizada pela divisão em um grande número de passos de tempo, o que implica uma granularidade fina na representação temporal. Por outro lado, uma série de baixa resolução é aquela que é dividida em um número menor de passos de tempo, indicando uma representação temporal mais generalizada. Uma série de eventos é considerada esparsa quando a quantidade de passos de tempo é substancialmente maior do que a quantidade de eventos observados.

Séries de eventos podem possuir o que se denomina na bibliografia de agrupamento temporal de eventos (ODENWELLER; DONNER, 2020). Eventos agrupados temporalmente ocorrem quando uma série esparsa possui dois ou mais eventos justapostos entre si (c.f. Figura 5). Formalmente, seja uma série de eventos x_i cuja quantidade total de eventos é s_i. Sejam l e $l+1$ índices para dois eventos subsequentes quaisquer dentro da série. Os eventos l e $l+1$ estão agrupados temporalmente se os instantes de tempo t_l^i e t_{l+1}^i, em que ocorrem l e $l+1$, respectivamente, diferem de uma unidade de tempo. Ou seja, dois eventos estão agrupados temporalmente se ocorrem em tempos subsequentes.

Figura 5 – Séries com padrão de agrupamento temporal



(a) Em séries com esse padrão, os eventos tendem a estar circundantes



(b) Todos os eventos desta série estão agrupados temporalmente

Séries ilustrativas geradas pelo autor

Existem algumas formas de medir de quão temporalmente agrupados os eventos estão. Uma delas é o *pairing coefficient*, ou coeficiente de emparelhamento (ODENWELLER; DONNER, 2020), presente na Equação 3.1:

$$P_i = \frac{1}{s_{i-1}} \sum_{l=1}^{s_i-1} \delta[(t_{l+1}^i - t_l^i) - 1] \tag{3.1}$$

Na Equação 3.1, P_i quantifica o agrupamento temporal de eventos em séries temporais e assume valores entre $P_i = 0$ (nenhum evento agrupado) e $P_i = 1$ (todos os eventos em passos de tempo subsequentes). A função $\delta(\cdot)$ assume o valor 1 apenas para argumento nulo, e 0 caso contrário. Além disso, nessa definição, a medida de tempo é adimensional. O coeficiente de emparelhamento, contudo, tem algumas limitações no tocante à sua interpretação no contexto do trabalho, como discutido na Seção 3.3.2.

3.2 Medidas de Similaridade

Existem muitas formas de medir semelhança, similaridade, relação, dependência ou associação entre variáveis. No contexto de séries temporais, muitas medidas diferentes foram utilizadas para quantificar essas associações, e, em termos gerais, essas medidas são nomeadas de medidas de similaridade. Cada medida tem sua empregabilidade determinada pela natureza das séries temporais (BOERS, 2015).

3.2.1 O Coeficiente de Correlação de Pearson e outras medidas

A medida de similaridade mais amplamente utilizada é o Coeficiente de Correlação de Pearson. Para duas séries temporais x e y de comprimento T com médias existentes $\overline{x}, \overline{y}$ e desvios padrão σ_x, σ_y, ela é definida como a forma bilinear na Equação 3.2:

$$\mathrm{Cor}(x,y) := \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma_{i=1}^{T}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{T}(x_i - \overline{x})^2 \sum_{i=1}^{T}(y_i - \overline{y})^2}} \tag{3.2}$$

Portanto, $Cor(x,y) \in [-1,+1]$ para todos os pares de x, y, sendo os valores próximos de 0 associados à baixa correlação, e os valores próximos a ± 1 indicam forte correlação (BENESTY *et al.*, 2009). Essa medida é adequada para quantificar as dependências lineares entre x e y. No entanto, isso não exclui possíveis dependências não lineares entre elas. A Figura 6 exemplifica pares com alta e baixa correlação.

A limitação dos coeficientes de correlação lineares impulsionou o desenvolvimento de medidas mais abrangentes de similaridade. Entre essas alternativas, destacam-se os coeficientes de correlação de postos de Spearman e Kendall, que quantificam dependências monótonas gerais entre x e y, incluindo não linearidades (KENDALL; GIBBONS, 1990).

O Coeficiente de Correlação de Spearman avalia a relação monotônica entre duas variáveis, valendo-se do conceito de postos. O conceito de postos está relacionado à atribuição de posições relativas a diferentes valores em um conjunto de dados. Quando se está lidando com duas variáveis, como no caso do Coeficiente de Correlação de Spearman, é comum classificar os valores de cada variável em ordem crescente, atribuindo a cada valor o seu "posto" na ordem (DANIEL, 2000). A correlação de Spearman é calculada por meio da Equação 3.3:

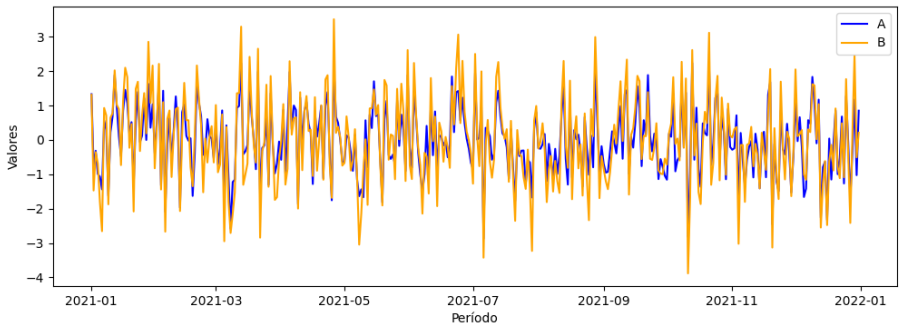
$$\mathrm{Cor}(x,y) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3.3}$$

em que d_i representa as diferenças entre os postos de cada par de observações (x_i, y_i), e n é o número total de observações. Assim como para Pearson, os valores de correlação de Spearman variam entre -1 (anticorrelação perfeita) e 1 (correlação perfeita), sendo 0 e seus valores circundantes um indicativo de ausência de correlação, vide a Figura 7.

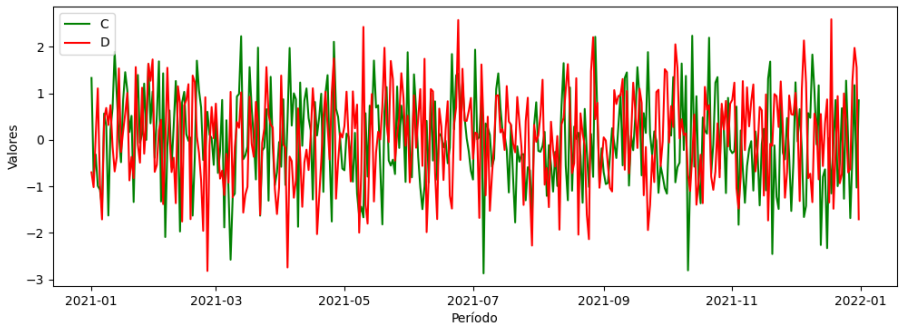
Já o Coeficiente de Kendall (KENDALL, 1938) mede a correlação entre pares de observações concordantes e discordantes, de acordo com a Equação 3.4:

$$\mathrm{Cor}(x,y) = \frac{\text{Número de pares concordantes} - \text{Número de pares discordantes}}{\frac{1}{2}n(n-1)} \tag{3.4}$$

Figura 6 – Pares de séries temporais com valores alto e baixo para Correlação de Pearson



(a) As séries acima têm Correlação de Pearson muito alta $\approx 0,921$



(b) Já o último par tem Correlação de Pearson baixa $\approx 0,146$

Séries ilustrativas geradas pelo autor

em que n é o número total de observações (c.f. Figura 8).

Mais detalhes acerca de medidas de similaridade aplicadas em séries temporais podem ser vistos em Wei (1994), Serrà e Arcos (2014), Keogh *et al.* (2001). A busca por medidas mais gerais, que transcendam dependências monótonas, levou ao desenvolvimento de conceitos como a Informação Mútua, uma medida de similaridade não linear que quantifica a "informação conjunta" contida em x e y. A Correlação de Informação Mútua (COVER; THOMAS, 2001) avalia a dependência estatística entre duas variáveis e pode ser calculada através da Equação 3.5:

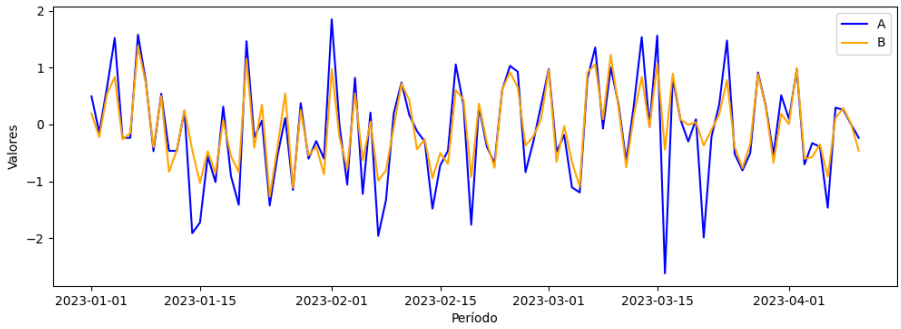
$$\mathrm{MI}(x, y) = \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i, y_j) \log \left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right) \tag{3.5}$$

em que $p(x_i, y_j)$ representa a probabilidade conjunta das observações x_i e y_j, e $p(x_i)$ e $p(y_j)$ são as probabilidades marginais de x_i e y_j. Essa medida proporciona uma avaliação robusta e abrangente da relação entre x e y, indo além das limitações impostas por métodos lineares. Ela atinge valores diversos a partir do conjunto de dados e sua unidade convencional é bits. A Figura 9 exemplifica uma aplicação.

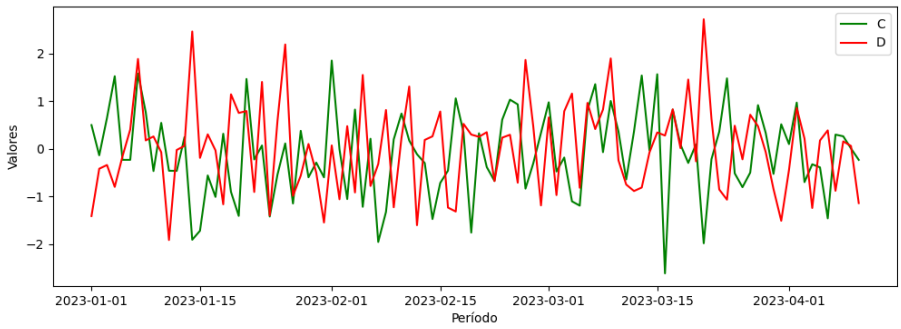
3.2.2 A limitação intrínseca ao problema

As medidas de similaridades apresentadas anteriormente possuem seu escopo de atuação e limitações. Em geral, todas as medidas propostas tendem a absorver muito bem correlação em

Figura 7 – Pares de séries temporais com valores alto e baixo para Correlação de Spearman



(a) Alta correlação de Spearman $\approx 0,941$



(b) Baixa correlação de Spearman $\approx -0,107$

Séries ilustrativas geradas pelo autor

contextos de variável contínua, pela própria natureza de suas definições. Contudo, em contextos discretos, especialmente binários, esse bom desempenho não é observado, como mostra o exemplo com as séries i e j presentes na Figura 10, em que a série j é definida a partir da série i, com seus eventos deslocados um passo de tempo para a direita. A evidente correlação entre essas séries não é bem captada pelas medidas anteriores, conforme pode-se inferir da Tabela 2.

Tabela 2 – Medidas de similaridade para as séries da Figura 10

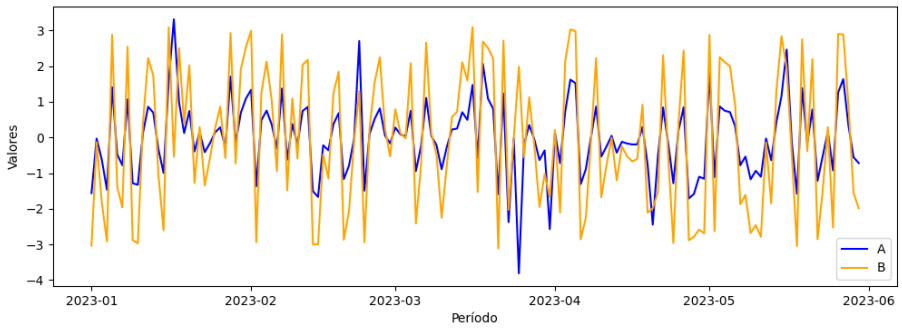
Medida de similaridade	**valor (três casas decimais)**
Pearson	0,114
Spearman	0,114
Kendall	0,114
Informação Mútua	0

Fonte: Elaborada pelo autor.

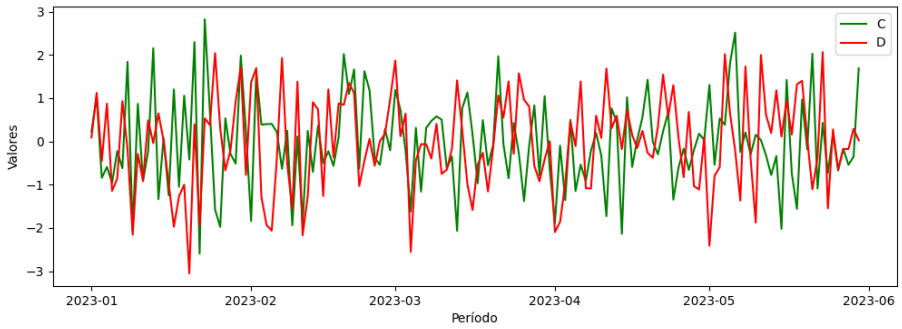
Os valores de correlação seriam ainda mais baixos quanto mais passos de tempo as duas séries possuíssem. Ademais, deslocar os dados em mais unidades de tempo para a direita ou esquerda não teria qualquer efeito significativo nessas correlações. O fato é que para dados binários, especialmente se forem esparsos, os padrões são difíceis de se estabelecer.

Para compreender plenamente o contexto do presente estudo, é crucial considerar a possibilidade de um atraso temporal entre eventos ocorridos em diferentes localidades, sendo esse

Figura 8 – Pares de séries temporais com valores alto e baixo para Correlação de Kendall



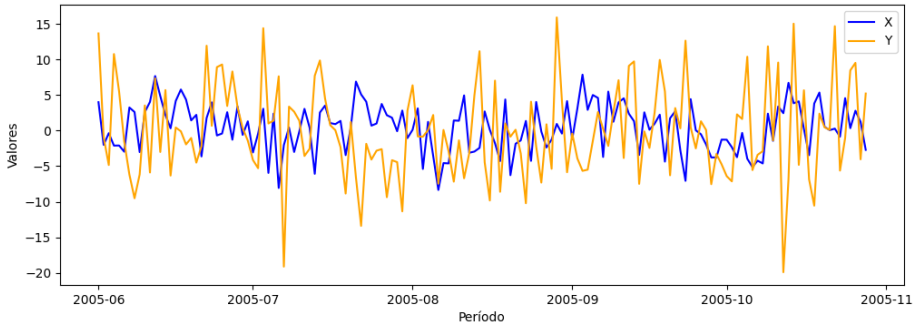
(a) Alta Correlação de Kendall $\approx 0,891$



(b) Baixa Correlação de Kendall $\approx 0,081$

Séries ilustrativas geradas pelo autor

Figura 9 – A Correlação de Informação Mútua entre X e Y é $\approx 0,029$

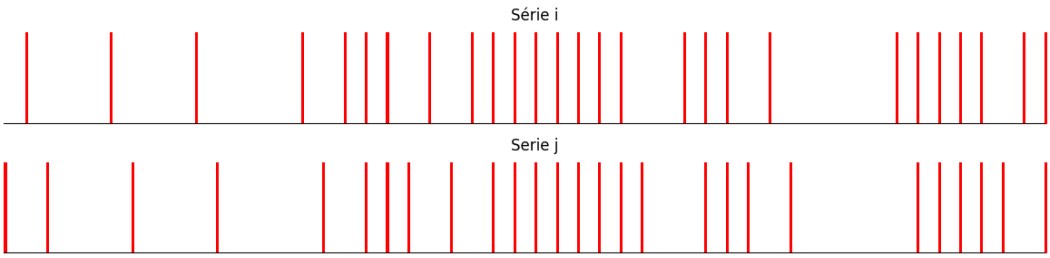


Séries ilustrativas geradas pelo autor

atraso não necessariamente constante ao longo do tempo. A existência de fatores interligando eventos criminais em locais distintos, como a presença de organizações criminosas, o impacto de medidas de segurança ou leis, ou mesmo fatores ocultos, pode resultar em padrões de ocorrências que se manifestam inicialmente em um local e, posteriormente, em outro, com variação nos períodos entre cada evento (RUITER, 2017; SLEEUWEN; STEENBEEK; RUITER, 2020).

Os atrasos entre os eventos em x e os eventos associados em y são influenciados por esses

Figura 10 – Séries de eventos com padrão muito similar



Séries ilustrativas geradas pelo autor

fatores e apresentam variações temporais. Para abordar essa dinâmica, é possível considerar deslocamentos temporais entre as séries temporais x e y por meio de janelas de tempo pré-definidas (avanço ou atraso), seguido pelo cálculo de medidas de similaridade. Entretanto, em uma "análise de avanço-atraso", apenas um único avanço (ou atraso) é atribuído ao par (x, y), assumindo-se como válido para todo o intervalo de tempo considerado.

Portanto, deve-se procurar uma medida de similaridade não linear que seja adequada para dados binários, forneça uma associação única entre eventos e permita um atraso dinâmico, ou seja, intervalos de tempo variáveis entre eventos de uma série temporal e eventos da outra série temporal. Uma opção possível para tal fim seria a *Event Synchronization* (ES), introduzida pela primeira vez em Quiroga, Kreuz e Grassberger (2002), que atende a todos esses requisitos.

3.3 Event Synchronization

O *Event Synchronization* foi introduzido como um método sem parâmetros para a análise de fenômenos de sincronização em dados espigados de eletroencefalografia, mas tem sido aplicado recentemente a outros campos de pesquisa também. A seguir, a definição original dada por Quiroga, Kreuz e Grassberger (2002) e sua respectiva correção:

Seja um conjunto de N séries temporais de eventos $\{x_1, ..., x_N\}$, cada uma de comprimento T, e (x_i, x_j) denota um par dessas séries temporais. Sejam dois eventos, l e m, pertencentes a duas séries distintas i e j, respectivamente, com quantidade de eventos s_i e s_j. Seja t_l^i e t_m^j o passo de tempo em que esses eventos ocorreram. Define-se a taxa de eventos r_i de uma série temporal x_i como o quociente entre o número de eventos s_i e o comprimento T de x_i: $r_i = \frac{s_i}{T}$.

Primeiramente, pode-se assumir que há uma taxa de evento característica, bem definida e igual para toda série temporal. Assim, pode-se permitir um atraso de tempo $\pm\tau^{ij}$ global entre eventos síncronos das séries i e j (que deve ser menor que a metade da distância mínima entre eventos, para evitar a contagem dupla). Denotando por $c(i|j)$ o número de vezes que um evento aparace em i e logo após aparecer em j, isto é,

$$c(i|j) = \sum_{l=1}^{s_i} \sum_{m=1}^{s_j} J_{lm}^{ij} \tag{3.6}$$

com

$$J_{lm}^{ij} = \begin{cases} 1 & \text{se } 0 < t_l^i - t_m^j \leq \tau^{ij} \\ \frac{1}{2} & \text{se } t_l^i = t_m^j \\ 0 & \text{caso contrário} \end{cases} \tag{3.7}$$

e de forma análoga para $c(j|i)$, define-se a combinação simétrica:

$$Q_{ij} = \frac{c(i|j) + c(j|i)}{\sqrt{s_i s_j}} \tag{3.8}$$

que mede a sincronização dos eventos entre as séries i e j e assume valores $Q_{ij} \in [0, 1]$. Tem-se $Q_{ij} = 1$ se e somente se os eventos estiverem totalmente sincronizados, e $Q_{ij} = 0$ na ausência de sincronização[2].

Nos casos em que se deseja evitar uma escala de tempo global τ^{ij} entre todos os eventos das séries, já que as taxas de eventos serão distintas em quase todas as séries temporais, usa-se a definição local τ_{lm}^{ij} para cada par de eventos l, m. Mais precisamente, define-se:

$$\tau_{lm}^{ij} = \frac{1}{2} \min \left(t_{l+1}^i - t_l^i, t_l^i - t_{l-1}^i, t_{m+1}^j - t_m^j, t_m^j - t_{m-1}^j \right) \tag{3.9}$$

e adequa-se J_{lm}^{ij} para a nova medida, com notação J_{lm} indicando intervalos locais (variáveis para cada l, m), conforme a Figura 11.

$$J_{lm}^{ij} = \begin{cases} 1 & \text{se } 0 < t_l^i - t_m^j \leq \tau_{lm}^{ij} \\ \frac{1}{2} & \text{se } t_l^i = t_m^j \\ 0 & \text{caso contrário} \end{cases} \tag{3.10}$$

Então:

$$c(i|j) = \sum_{l=1}^{s_i} \sum_{m=1}^{s_j} J_{lm}^{ij} \tag{3.11}$$

e $Q_{ij} = \dfrac{c(i|j) + c(j|i)}{\sqrt{s_i s_j}}$, como anteriormente.

O fator $\frac{1}{2}$ na definição de τ_{lm}^{ij} evita a contagem dupla se, por exemplo, dois eventos da série i estão próximos a um mesmo evento em j. Em alguns casos, especialmente para séries esparsas, é possível que τ_{lm}^{ij}, como definido anteriormente, torne-se indesejadamente grande, o que pode ser evitado a partir de um parâmetro $\tau_{\max}$, definido previamente à aplicação de ES.

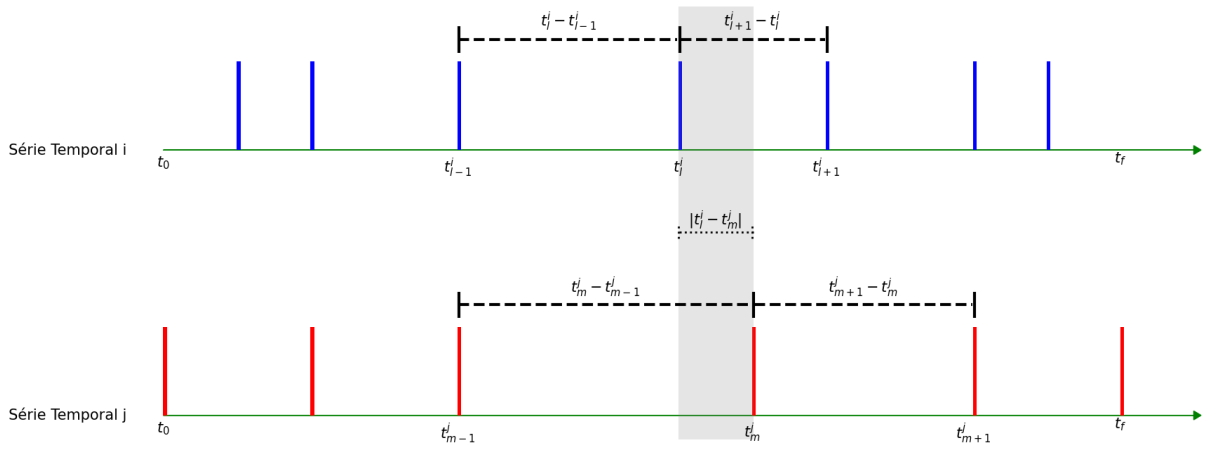
A definição anterior é bastante flexível e consegue absorver bem a natureza do problema de medir similaridade entre séries de eventos, mas ainda precisa de adaptações que delimitem melhor a análise e corrijam a contagem múltipla de eventos. Elas foram propostas por Odenweller e Donner (2020):

Dois eventos em t_l^i e t_m^j são considerados sincronizados se ambos ocorrerem dentro de um certo intervalo de tempo adaptativo aos dados, com largura τ_{lm}^{ij} definida como na Equação 3.12:

$$\tau_{lm}^{ij} = \frac{1}{2} \min \left(t_{l+1}^i - t_l^i, t_l^i - t_{l-1}^i, t_{m+1}^j - t_m^j, t_m^j - t_{m-1}^j \right) \tag{3.12}$$

[2] Duas séries com valor suficientemente grande de ES são ditas sincronizadas em vez de correlacionadas.

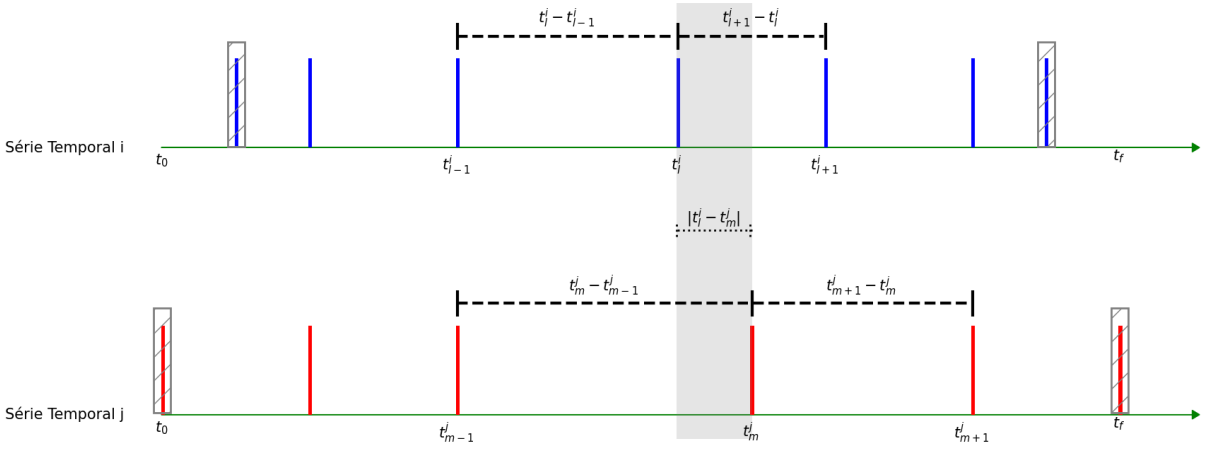
Figura 11 – Ilustração da análise de padrões feita no ES



Gerada pelo autor

com $l = 2, 3, ..., s_i - 1$ e $m = 2, 3, ..., s_j - 1$, de modo que τ_{lm}^{ij} não seja avaliado para o primeiro e último evento, a fim de garantir uma consideração apropriada das fronteiras (c.f. Figura 12). Portanto, não se computa τ_{1m} ou τ_{l1}, nem $\tau_{s_i m}$ ou $\tau_{l s_j}$. Ainda sim, estes são usados para o segundo e penúltimo evento (como t_{l-1}^i ou t_{l+1}^i). Novamente, é possível limitar superiormente τ_{lm}^{ij} a partir de $\tau_{\max}$, para evitar intervalos grandes demais entre eventos.

Figura 12 – Análise de padrões com ES corrigida



Gerada pelo autor

A Equação 3.12 implica que quanto mais raramente os eventos ocorrem em uma ou ambas as séries temporais, maior será τ_{lm}^{ij}, de modo que ela se configura como um intervalo de coincidência dinâmica (local). Assim, se os eventos são raros nas proximidades de um dos dois eventos, maiores desvios de uma coincidência instantânea ainda podem ser considerados sincronizados. A natureza dinâmica de τ_{lm}^{ij} simplifica a separação de eventos independentes, o que, por sua vez, resulta em uma variedade de escalas temporais capturadas por uma única

medida. A compensação é que, por design, o valor de τ_{lm}^{ij} muda constantemente entre diferentes pares de eventos.

Contar o número de ocorrências de eventos sincronizados em i, dado um evento em j, resulta em

$$c(i|j) = \sum_{l=2}^{s_i-1} \sum_{m=2}^{s_j-1} J_{lm}^{ij} \tag{3.13}$$

em que J_{lm}^{ij} é uma função de contagem que incorpora τ_{lm}^{ij} e depende se a condição de sincronização presente na Equação 3.14:

$$\sigma_{lm}^{ij} = \begin{cases} 1 & \text{se } 0 < t_l^i - t_m^j \leq \tau_{lm}^{ij} \\ 0 & \text{caso contrário} \end{cases} \tag{3.14}$$

é satisfeita para os eventos considerados e vizinhos:

$$J_{lm}^{ij} = \begin{cases} 1 & \text{se } \sigma_{lm}^{ij} = 1 \text{ e } \sigma_{m,l-1}^{ji} = 0 \text{ e } \sigma_{m+1,l}^{ji} = 0, \\ \frac{1}{2} & \text{se } t_l^i = t_m^j \text{ ou } (\sigma_{lm}^{ij} = 1 \text{ e } (\sigma_{m,l-1}^{ji} = 1 \text{ ou } \sigma_{m+1,l}^{ji} = 1)), \\ 0 & \text{caso contrário.} \end{cases} \tag{3.15}$$

A função de contagem na equação anterior difere da definição original de ES e essas mudanças são inevitáveis para uma especificação correta, pois, do contrário, a contagem dupla errônea poderia ocorrer. Devido à condição de uma distância entre eventos que é menor ou igual ao intervalo de coincidência dinâmica τ_{lm}^{ij}, na definição original os eventos poderiam ser contados duas vezes. Para evitar isso, é preciso verificar para todos os pares de eventos se um dos eventos já foi contado como sincronizado na direção oposta. Se for o caso, um peso de $\frac{1}{2}$ é atribuído a esse par, garantindo assim que a normalização seja feita corretamente. Essa situação só pode ocorrer se $t_l^i - t_m^j = \tau_{lm}^{ij}$ e, em seguida, os eventos respectivos contribuem igualmente para $c(i|j)$ e $c(j|i)$.

Por plena analogia, define-se ainda $c(j|i)$ e infere-se a força de sincronização de eventos entre i e j como:

$$Q_{ij}^{sym} = \frac{c(i|j) + c(j|i)}{\sqrt{(s_i - 2)(s_j - 2)}} \tag{3.16}$$

que é normalizada, de modo que $0 \leq Q_{ij}^{sym} \leq 1$, onde $Q_{ij}^{sym} = 1$ implica uma sincronização completa de eventos e $Q_{ij}^{sym} = 0$ a ausência de eventos sincronizados.

Para a geração de uma representação de rede complexa de um conjunto de séries temporais, considera-se a força de sincronização de eventos como uma medida estatística de similaridade, cujos valores estimados fornecem os coeficientes de uma matriz $\mathcal{Q}^{sym} = (Q_{ij}^{sym})$. Uma vez que Q_{ij}^{sym}, conforme definido acima, é simétrico em relação a permutações entre i e j, essa matriz é simétrica e pode, portanto, ser usada para construir uma rede não direcionada a partir de dados de eventos. Contudo, essa definição simétrica para a matriz pode ter algumas limitações (vide Seção 3.3.1).

Dadas duas séries temporais x_i e x_j, medidas nos locais i e j, pode-se estar interessado no número total de eventos síncronos que ocorreram primeiro em j e depois em i, e, separadamente, no número total de eventos síncronos que ocorreram primeiro em i e depois em j. Por esse

motivo, uma versão modificada da *Event Synchronization* direcionada foi introduzida (BOERS *et al.*, 2014), onde as somas correspondentes serão armazenadas separadamente. Além disso, especialmente ao aplicar ES a dados com alta resolução temporal, podem ocorrer situações em que vários eventos ocorrem durante passos de tempo consecutivos (agrupamento temporal). Nessas situações, apenas o primeiro será considerado como um evento, ponderado pelo número de eventos subsequentes, que são descartados da soma. Portanto, para cada evento l em i , há um peso w_l^i. Então, define-se W_{lm}^{ij} como:

$$W_{lm}^{ij} = \begin{cases} \min(w_l^i, w_m^j) & \text{se } 0 < t_l^i - t_m^j \leq \tau \text{ e } t_l^i - t_m^j \leq \tau_{\max}, \\ 0 & \text{caso contrário.} \end{cases} \tag{3.17}$$

em que deve ser enfatizado que eventos exatamente no mesmo tempo não contribuem, pois não permitem determinar a ordem temporal. A introdução de pesos w_l^i acima assegura que em situações em que há agrupamento temporal em uma determinada série de eventos x_j, seguido por um agrupamento temporal em outra série x_i, com sobreposição temporal entre os dois, todos os eventos ainda são contados de maneira ordenada no tempo.

Define-se *Event Synchronization* direcionada como:

$$Q_{ij}^{\text{dir}} = \frac{\sum_{l=2}^{s_i-1} \sum_{m=2}^{s_j-1} W_{lm}^{ij}}{\sqrt{(s_i - 2) \cdot (s_j - 2)}} \tag{3.18}$$

que não é necessariamente simétrico: em geral, $Q_{ij}^{\text{dir}} \neq Q_{ji}^{\text{dir}}$.

Para gerar redes complexas a partir dessa medida, considera-se a força de sincronização de eventos como uma medida estatística de associação genérica, cujos valores estimados fornecem os coeficientes de uma matriz $\mathcal{Q}^{dir} = (Q_{ij}^{dir})$. Uma vez que Q_{ij}^{dir} não é simétrico em relação a permutações entre i e j, $\mathcal{Q}^{dir}$ não é simétrica e pode, portanto, ser usada para construir uma rede direcionada.

3.3.1 A taxa de eventos

Apesar do fator de normalização $\sqrt{(s_i - 2) \cdot (s_j - 2)}^{-1}$ nas equações , o valor de ES depende das taxas de eventos r_i e r_j se $\tau_{\max}$ for finito, já que a probabilidade de sincronizações "aleatórias" aumenta com o aumento das taxas de eventos. Os valores da matriz $\mathcal{Q}$, $Q_{ij} \in [0, 1]^{N \times N}$, calculados para diferentes pares de séries de eventos não são diretamente comparáveis se a taxa de eventos variar entre as séries de eventos (BOERS, 2015).

A maneira mais intuitiva de obter valores comparáveis com ES é definir o conceito de evento de modo que a taxa de eventos seja igual para todas as séries em consideração (BOERS, 2015). No entanto, algumas definições comuns, por construção, não permitem taxas de eventos iguais em todas as séries de eventos (por exemplo, se os eventos são definidos como passos de tempo nos quais os valores correspondentes estão acima de um limiar global), exigindo assim uma solução mais sofisticada para esse problema.

Uma abordagem adequada em tais situações é comparar os valores de Q_{ij} em termos de sua significância estatística, que por sua vez depende das taxas de eventos r_i e r_j das séries de eventos consideradas x_i e x_j. Modelos estatísticos nulos apropriados para ES dependem da

definição específica de eventos. Denotando a função de densidade de probabilidade correspondente por H_{r_i,r_j}, a significância estatística de um dado valor empírico Q_{ij} pode ser estimada pela probabilidade de obter um valor $\mathfrak{Q}_{ij}$, maior ou igual a Q_{ij}:

$$P(\mathfrak{Q}_{ij} \geq Q_{ij}) = \int_{Q_{ij}}^{1} H_{r_i,r_j}(s)\, ds \tag{3.19}$$

Ao contrário dos valores Q_{ij} em si, suas posições na distribuição do modelo $P(\mathfrak{Q}_{ij} \geq Q_{ij})$ são comparáveis entre pares de séries de eventos com diferentes taxas de eventos (r_i, r_j). Nota-se que nesta abordagem, é possível omitir completamente a normalização por $\sqrt{s_i \cdot s_j}$ nas equações anteriores (BOERS, 2015).

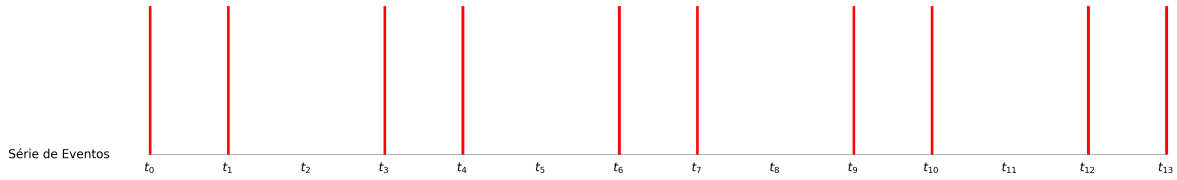
3.3.2 O agrupamento temporal nas séries de eventos

Em séries temporais de eventos, os eventos podem estar potencialmente agrupados temporalmente, o que gera um viés ao aplicar a *Event Synchronization* (ES) em sua definição original. Esse viés surge principalmente da redução do intervalo de coincidência dinâmica local para $\frac{1}{2}$ em cálculos de pares eventos agrupados, permitindo apenas a coincidência simultânea (ODENWELLER; DONNER, 2020).

Para o *Event Synchronization* direcionado, a definição já lida com esse problema, a partir do conceito de pesos. Mas para a rede simétrica, o agrupamento temporal faz com que a quantidade de conexões seja subestimada (ODENWELLER; DONNER, 2020), pois nos eventos em que ele ocorre, apenas coincidências simultâneas influenciariam no cálculo da similaridade.

Uma maneira de saber quão agrupados temporalmente os dados estão foi indicada na Seção 3.1.3, o coeficiente de emparelhamento, mas ele sofre de uma limitação de interpretação, uma vez que não há uma faixa de valores para quantificá-lo como alto ou baixo. Por exemplo, na figura 13, vê-se uma série cujos eventos sofrem todos com o agrupamento temporal. Contudo, seu coeficiente de emparelhamento é aproximadamente $0,57$.

Figura 13 – Todos os eventos ocorrem em pares



Série ilustrativa gerada pelo autor

Assim, uma medida que aparenta ser mais interpretável é aqui proposta, como ”taxa de agrupamentos”, cuja ideia é contar quantos eventos agrupados (que estão em um ’grupo’, em que grupo é uma coleção de dois ou mais eventos subsequentes) existem em relação ao total. A medida expressa a porcentagem de eventos que estão antecedidos ou sucedidos por outros, pois é isso que faz o intervalo de coincidência local ir para $\frac{1}{2}$:

$$R_i = \frac{1}{s_i}\left(\delta[t_2^i - t_1^i - 1] + \delta[t_{s_i}^i - t_{s_i-1}^i - 1] + \sum_{l=2}^{s_i-1} \delta[(t_l^i - t_{l-1}^i - 1)\cdot(t_{l+1}^i - t_l^i - 1)]\right) \quad (3.20)$$

A primeira parcela dentro dos parênteses, $\delta[t_2^i - t_1^i - 1]$, verifica se o primeiro evento da série está agrupado temporalmente (com o segundo). A segunda parcela, $\delta[t_{s_i}^i - t_{s_i-1}^i - 1]$, verifica se o último evento está agrupado temporalmente (com o penúltimo). O somatório verifica evento a evento, do segundo ao penúltimo, se ele está agrupado temporalmente com seu antecessor ou sucessor, a partir de $(t_l^i - t_{l-1}^i - 1)$ e $(t_{l+1}^i - t_l^i - 1)$, respectivamente. Se estiver, o parênteses vai a zero, assim como o produto, e este evento é somado. Assim, há uma análise evento a evento, para encontrar aqueles a um passo de tempo de seus adjacentes.

3.4 Redes Complexas

O termo redes complexas é relativamente recente. Ele começou a ser usado no final da década de 1990, quando pesquisadores de disciplinas muito distintas - cientistas da computação, biólogos, sociólogos, físicos e matemáticos - passaram a estudar de forma intensiva as redes do mundo real e seus modelos. As limitações das metodologias de investigação da época para alguns contextos proporcionaram o desenvolvimento da teoria subsequente (ALBERT; BARABáSI, 2002; DOROGOVTSEV, 2010).

Em termos muito gerais, uma rede é qualquer sistema que admita uma representação matemática abstrata como um grafo, cujos nós (vértices) identificam os elementos do sistema e em que o conjunto de conexões (arestas) representa a presença de uma relação ou interação entre esses elementos. Claramente, esse alto nível de abstração se aplica a uma ampla gama de sistemas. Nesse sentido, as redes fornecem um arcabouço teórico que permite uma representação conceitual conveniente de interconexões em sistemas complexos, nos quais a caracterização do nível do sistema implica o mapeamento das interações entre um grande número de indivíduos (BARRAT, 2013).

Com o aumento significativo do poder computacional nos últimos anos, o estudo de sistemas interconectados em grande escala tem experimentado um avanço notável. Isso se deve, em grande parte, à disponibilidade crescente de conjuntos de dados extensos e à capacidade dos computadores para armazenar e manipular essas informações de forma eficiente. À medida que a capacidade de processamento de dados continua a avançar e as técnicas de análise de dados se tornam cada vez mais sofisticadas, é natural esperar um crescimento constante e um desenvolvimento exponencial nesta área de pesquisa (BARRAT, 2013).

3.4.1 Definições formais

Matematicamente, uma rede é representada como um grafo $\mathcal{G}(V, E)$, isto é, um objeto que consiste em um conjunto de nós (ou vértices) V representando os objetos (ou agentes) na rede, e um conjunto E de arestas (ou conexões) representando as interações ou relações entre os nós (c.f. Figuras 14 e 15). A cardinalidade desses conjuntos, que representa o número total de nós e arestas, é geralmente denotada por N e M, respectivamente. As arestas podem ser direcionadas ou não direcionadas, com ou sem pesos.

Figura 14 – Rede com nós representando personagens de Les Misérables

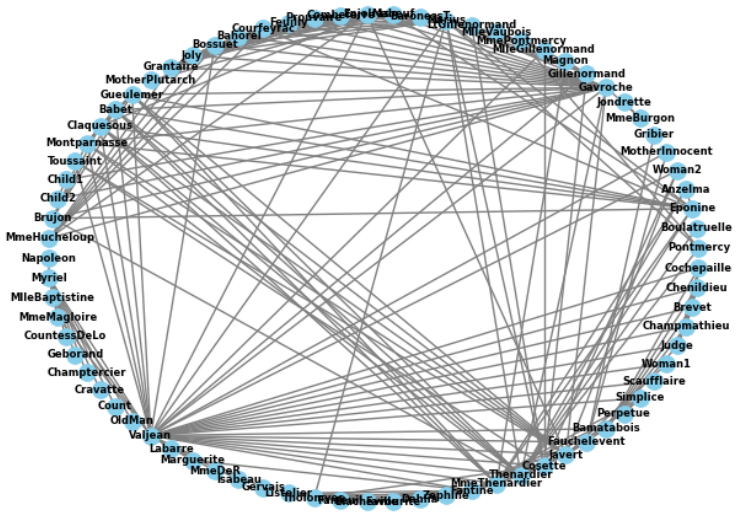


Figura 15 – Outra possibilidade de disposição para a mesma rede. Percebe-se que a visualização de uma grande rede é uma tarefa complexa

Figuras ilustrativas geradas pelo autor

O escopo do presente trabalho não contempla grafos $\mathcal{G}$ que possuem mais de uma aresta para um mesmo par de nós (multigrafo), ou que possuem arestas que partem de um nó para ele mesmo (laços). Esses tipos de grafo possuem uma modelagem muito própria que não é contemplada pelo que será visto a seguir.

A topologia de uma rede é muitas vezes representada pela matriz de adjacências $A_{N\times N}$, em que o elemento a_{ij} representa uma aresta entre o nó i e o nó j. No caso de um grafo não direcionado e sem pesos, os elementos a_{ij} da matriz de adjacências A assumem valores $a_{ij} = a_{ji} = 1$ se o nó i e o nó j estão conectados e $a_{ij} = a_{ji} = 0$ se não estão. Nesse caso, o não direcionamento do grafo torna sua matriz de adjacências simétrica, explicitando uma relação

bidirecional, e a falta de pesos significa que não há hierarquia entre as conexões estabelecidos pelos diferentes nós. Matrizes ilustrativas são apresentadas nas Figuras 16, 17, 18, e 19.

Figura 16 – Matriz de adjacências para grafo não direcionado sem pesos

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Figura 17 – Matriz de adjacências para grafo direcionado sem pesos

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

Figura 18 – Matriz de adjacências para grafo não direcionado com pesos

$$\begin{bmatrix} 0 & 2 & 0 & 4 \\ 2 & 0 & 5 & 1 \\ 0 & 5 & 0 & 0 \\ 4 & 1 & 0 & 0 \end{bmatrix}$$

Figura 19 – Matriz de adjacências para grafo direcionado com pesos

$$\begin{bmatrix} 0 & 3 & 0 & 7 \\ 0 & 0 & 1 & 0 \\ 0 & 3 & 0 & 0 \\ 1 & 0 & 4 & 0 \end{bmatrix}$$

Exemplos ilustrativos gerados pelo autor

Algumas redes representam relações unidirecionais entre os nós que não são perfeitamente capturadas por arestas bidirecionais. Nesse caso, a aresta a_{ij} representa uma conexão que sai do nó i e vai para o nó j, e não necessariamente assume o mesmo valor que a_{ji} (que sai do nó j para o nó i). A matriz de adjacências, portanto, deixa de ser simétrica, e a rede é dita direcionada.

Pesos em arestas de um grafo geralmente são motivados por diferenças sistemáticas entre a importância e a natureza das conexões entre dois nós. Quando a modelagem os considera, a matriz A tem elementos a_{ij} que podem assumir, a princípio, qualquer valor. A ligação agora deixa de ter um significado booleano entre existir e não existir e passa a representar algum tipo de grandeza, e a rede torna-se ponderada. Para o presente trabalho, são de interesse as redes direcionadas e não direcionadas, sem pesos.

3.4.2 Medidas estruturais

Em uma rede com N nós, o número máximo de arestas é dado por $M_{max} = N \cdot (N-1)$. Contudo, na maioria das redes do mundo real, apenas uma pequena parte das possíveis arestas é deferida. Para avaliar quão conectada é a rede, define-se uma medida que representa a razão entre a quantidade de arestas existentes e a máxima possível, chamada de Densidade de Conexões (DONNER; WIEDERMANN; DONGES, 2017):

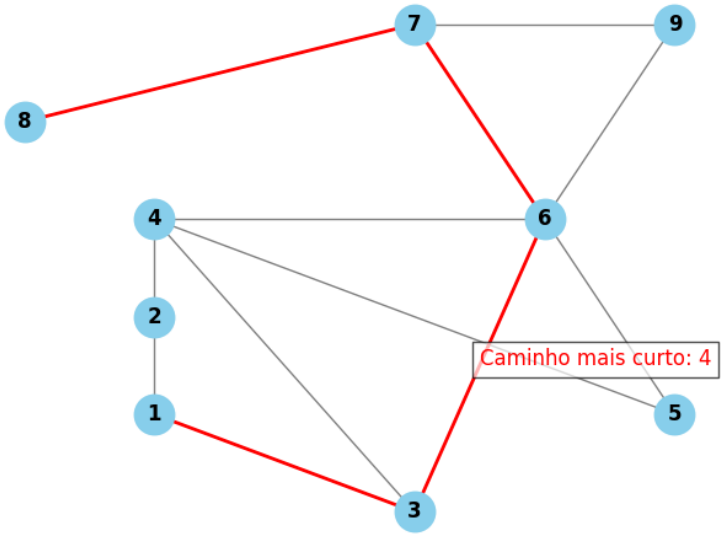
$$\rho = \frac{2M}{M_{max}} = \frac{\sum_{ij} a_{ij}}{N(N-1)} \tag{3.21}$$

As arestas existentes entre os nós acabam por definir caminhos, que levam de um determinado nó até outro. Uma maneira de entender o quão interconectados os nós de uma rede estão é calcular o comprimento médio do caminho mais curto entre dois nós (NEWMAN, 2010):

$$\langle l(i,j) \rangle = \frac{1}{N(N-1)} \sum_{i \neq j}^{N} l(i,j) \tag{3.22}$$

Na Equação 3.22, $l(i,j)$ mede o comprimento do caminho mais curto entre o nó i e o nó j, que denota o número de arestas que existe no menor caminho entre eles (c.f. Figura 20). Nota-se que para calcular o comprimento médio do caminho mais curto, é necessário lidar com um grafo conexo.

Figura 20 – Exemplo de caminho mais curto entre os nós 1 e 8, com comprimento igual a 4



Rede ilustrativa gerada pelo autor

A excentricidade em uma rede refere-se ao maior comprimento do caminho mais curto entre um nó específico e todos os outros nós da rede (HAGE; HARARY, 1995). Em outras palavras, a excentricidade de um nó é o número máximo de arestas que deve percorrer para alcançar o nó mais distante a partir dele. Formalmente, a excentricidade e_i de um nó i é dada por:

$$e_i = \max_{j \neq i} l(i,j) \tag{3.23}$$

em que $l(i,j)$ é o comprimento do caminho mais curto entre os nós i e j. A excentricidade fornece uma medida da "distância máxima" de um nó para os demais na rede. Ela é crucial para o cálculo de outras métricas topológicas, como o raio da rede, que é o menor valor das excentricidades, representado por:

$$R = \min_{i} e_i \tag{3.24}$$

Além disso, a excentricidade também é utilizada no cálculo do diâmetro da rede, que é o maior valor das excentricidades:

$$D = \max_i e_i \tag{3.25}$$

3.4.3 Medidas globais

Pode-se quantificar a existência de conexões transitivas em uma rede, calculando sua transitividade T. Ela é a razão entre o número de triângulos na rede e o número de ternas conectadas (NEWMAN, 2010), podendo ser calculada pela matriz de adjacências. A fórmula para a transitividade T é dada por:

$$T = \frac{\sum_{i,j,k=1}^{N} a_{ij}a_{ik}a_{jk}}{\sum_{i,j,k=1,j\neq k}^{N} a_{ij}a_{ik}} \tag{3.26}$$

Na Equação 3.26, o numerador representa o número de triângulos na rede, pois $a_{ij}a_{ik}a_{jk}$ é igual a 1 somente quando há arestas entre todos os três nós, formando um triângulo. O denominador representa o número total de ternas conectadas, excluindo as ternas em que j é igual a k, para evitar a contagem de laços e de conexões duplicadas. Assim, a transitividade T oferece uma medida da propensão da rede em formar triângulos, indicando o quão frequentemente os vizinhos de um nó estão conectados entre si. Redes mais transitivas tendem a exibir padrões de conexões mais fortemente interligados e organizados.

Assortatividade (NEWMAN, 2010) em uma rede refere-se à tendência dos nós de se conectarem a outros nós que possuem características semelhantes. Essa propriedade pode ser observada através de diferentes atributos dos nós, como grau (número de conexões) ou outros traços específicos. A medida, muitas vezes denotada por r, quantifica essa preferência de conexão entre nós similares. Para redes baseadas no grau dos nós, a assortatividade pode ser calculada pela correlação entre os graus dos nós conectados. A fórmula geral para r é dada por:

$$r = \frac{\sum_{jk} jk(m_{jk} - q_j q_k)}{\sigma_q^2} \tag{3.27}$$

em que:
- j e k representam os graus dos nós conectados,
- m_{jk} é o número de arestas entre nós de grau j e k,
- q_j e q_k são as frações de arestas ligadas a nós de grau j e k, respectivamente, em relação ao total de arestas,
- σ_q^2 é a variância de q_j.

Se $r > 0$, a rede é assortativa, indicando uma preferência por conexões entre nós de graus similares. Se $r < 0$, a rede é dissortativa, sugerindo uma tendência para conexões entre nós de graus diferentes.

3.4.4 Medidas locais

Para investigar o papel e a importância dos nós em uma rede, diversas medidas de centralidade foram propostas, que frequentemente levam em consideração propriedades muito

específicas. Elas quantificam diferentes aspectos da posição de um nó, indicando a sua relevância para a estrutura da rede.

O grau do nó é a medida mais básica dentre todas as medidas de centralidade e simplesmente equivale ao número de arestas adjacentes a um único nó (NEWMAN, 2010). Isso pode ser alcançado matematicamente por uma soma sobre as colunas da matriz de adjacência. Em uma rede não direcionada e não ponderada, o grau k do nó i é dado por:

$$k_i = \sum_{j=1}^{N} a_{ij} \tag{3.28}$$

Seguindo uma lógica diferente, podemos medir a centralidade de um nó calculando quantos caminhos mais curtos em uma rede passam por ele. A medida de centralidade correspondente, intermediância b_i, avalia a importância de um nó i com base na frequência com que ele atua como ponte ao longo dos caminhos mais curtos entre outros nós na rede (NEWMAN, 2010). Ela é definida por:

$$b_i = \sum_{j \neq i \neq k} \frac{l(j,k|i)}{l(j,k)} \tag{3.29}$$

em que i, j e k servem como rótulos de nó e $l(j,k|i)$ denota um caminho mais curto entre o nó j e k que passa pelo nó i. Assim, intermediância é o número total de caminhos mais curtos entre j para k que passam pelo nó i dividido pelo total de caminhos entre j e k. Um nó que possui muitos caminhos mais curtos que passam por si possui uma intermediância alta.

Para quantificar a conectividade da vizinhança de um nó, foi introduzido o coeficiente de agrupamento local c (NEWMAN, 2010):

$$c_i = \frac{2m_i}{k_i(k_i - 1)} \tag{3.30}$$

Na Equação 3.30, m_i representa o número de arestas que conectam os vizinhos topológicos do nó i. O coeficiente de agrupamento local assume valores entre 0 (nenhum nó vizinho está conectado) e 1 (vizinhos formam um subgrafo completo).

3.4.5 Medidas em redes direcionadas

Em redes direcionadas, as medidas estruturais consideram a orientação das arestas, adicionando complexidade à análise topológica (NEWMAN, 2010). Para calcular a densidade de conexões em uma rede direcionada, é preciso distinguir entre arestas de chegada e saída entre os nós. A densidade de conexões (ρ) em uma rede direcionada com N nós é dada por:

$$\rho = \frac{M}{N(N-1)} = \frac{\sum_{ij} a_{ij}}{N(N-1)} \tag{3.31}$$

em que M é o número total de arestas na rede. No caso de redes direcionadas, a matriz A não é mais simétrica, e $\sum_{ij} a_{ij}$ agora conta individualmente cada aresta, devido à orientação das arestas.

O comprimento médio do caminho mais curto entre dois nós ($\langle l(i,j) \rangle$) em uma rede direcionada leva em consideração a direção das arestas e pode ser definido como:

$$\langle l(i,j) \rangle = \frac{1}{N(N-1)} \sum_{i \neq j}^{N} l(i,j) \tag{3.32}$$

em que $l(i, j)$ representa o comprimento do caminho mais curto da forma mais direta possível de i para j.

A excentricidade (e_i) em uma rede direcionada refere-se novamente ao comprimento máximo do caminho mais curto de um nó específico para todos os outros nós na rede. O raio (R) e o diâmetro (D) da rede direcionada são análogos aos conceitos em redes não direcionadas.

$$R = \min_i e_i \tag{3.33}$$

$$D = \max_i e_i \tag{3.34}$$

3.4.6 Medidas globais em Redes Direcionadas

A transitividade (T) em redes direcionadas leva em conta a formação de triângulos, considerando entra. A fórmula para a transitividade (T) em uma rede direcionada é ajustada para refletir a orientação das arestas:

$$T = \frac{\sum_{i,j,k=1}^{N} a_{ji}a_{ik}a_{kj}}{\sum_{i,j,k=1,j\neq k}^{N} a_{ji}a_{ki}} \tag{3.35}$$

A assortatividade em redes direcionadas pode ser analisada considerando a correlação entre os graus dos nós conectados. A fórmula geral para (r) em redes direcionadas pode levar em conta o grau de saída, de entrada, ou a soma:

$$r = \frac{\sum_{jk} jk(m_{jk} - q_j q_k)}{\sigma_q^2} \tag{3.36}$$

Uma nova métrica que se faz útil no contexto de redes direcionadas é a reciprocidade (GARLASCHELLI; LOFFREDO, 2004). A reciprocidade é uma medida que quantifica a tendência das relações entre os nós a serem "bidirecionais", é a razão entre o número de arestas recíprocas (quando dois nós i,j possuem arestas tanto de i para j quanto de j para i) pelo total de arestas.

A reciprocidade pode ser expressa considerando a presença de arestas em ambas as direções. Se a_{ij} representa a existência de uma aresta direcionada de i para j e a_{ji} representa a existência da aresta de j para i, então a reciprocidade pode ser calculada como:

$$R = \frac{\sum_{i,j} a_{ij} \cdot a_{ji}}{\sum_{i,j} a_{ij}} \tag{3.37}$$

A Equação 3.37 compara o número de arestas bidirecionais com o total de arestas na rede, fornecendo uma medida relativa de reciprocidade. Em uma rede altamente recíproca, os nós têm uma propensão significativa para se relacionarem reciprocamente, enquanto em uma rede com baixa reciprocidade, as relações são predominantemente unidirecionais.

3.4.7 Medidas locais em Redes Direcionadas

A medida de grau (k_i) em uma rede direcionada representa o número de arestas que saem ou entram em um nó específico:

$$k_i^{out} = \sum_{j=1}^{N} a_{ij} \tag{3.38}$$

$$k_i^{in} = \sum_{j=1}^{N} a_{ji} \tag{3.39}$$

Dessa forma, muitas medidas de rede que usam o conceito de grau podem ser estendidas. Uma medida que é apenas definida para redes direcionadas é a divergência de rede Δk (WOLF, 2021). Calcula-se a divergência de rede considerando a diferença entre os graus de entrada e saída:

$$\Delta k_i = k_{\text{in},i} - k_{\text{out},i} \tag{3.40}$$

e interpreta-se a divergência de rede como uma indicação de fontes e sumidouros em uma rede.

A intermediância (b_i) em uma rede direcionada é calculada levando em conta a frequência com que um nó atua como ponte ao longo dos caminhos mais curtos entre outros nós, considerando a direção das arestas:

$$b_i = \sum_{j \neq i \neq k} \frac{l(j,k|i)}{l(j,k)} \tag{3.41}$$

O coeficiente de agrupamento local (c_i) em uma rede direcionada também leva em consideração a orientação das arestas:

$$c_i^{out} = \frac{\sum_{j,k=1}^{N} a_{ij} a_{ik} a_{jk}}{k_i^{out}(k_i^{out} - 1)} \tag{3.42}$$

$$c_i^{in} = \frac{\sum_{j,k=1}^{N} a_{ji} a_{ki} a_{jk}}{k_i^{in}(k_i^{in} - 1)} \tag{3.43}$$

Essas extensões oferecem uma compreensão mais refinada das propriedades estruturais, globais e locais, levando em consideração a direcionalidade na análise.

3.4.8 Componentes conexas e Comunidades

Componentes conexas em um grafo são conjuntos de nós que estão interligados por arestas, formando subgrafos onde cada par de nós está conectado por pelo menos um caminho (BARABáSI; POSFAI, 2016; NEWMAN, 2010). Em outras palavras, um componente conexo é um subconjunto do grafo original no qual é possível chegar de qualquer nó para qualquer outro nó por meio de arestas, seguindo um caminho na subestrutura.

Formalmente, um componente conexo C_i é definido como um subgrafo $G_i = (V_i, E_i)$, em que:
- V_i é o conjunto de nós pertencentes ao componente conexo,
- E_i é o conjunto de arestas que conectam os nós em V_i,
- Para cada par de nós i, j em V_i, existe pelo menos um caminho em G_i que os conecta.

Grafos não direcionados podem ter um ou mais componentes conexas, enquanto em grafos direcionados, é possível falar em componentes fortemente conexas (onde há caminhos em ambas as direções entre quaisquer dois nós) ou fracamente conexas (onde considera-se a direção das arestas) (NAGAMOCHI; IBARAKI, 2008). A identificação de componentes conexas é uma parte fundamental da análise de grafos, pois fornece pistas sobre a estrutura global, destacando

grupos de nós que estão mais intimamente ligados entre si em comparação com o restante do grafo (NAGAMOCHI; IBARAKI, 2008).

Para analisar uma rede em escala mesoscópica, o conceito de comunidades foi desenvolvido (WOLF, 2021). Uma comunidade dentro de uma rede descreve um conjunto de nós altamente interconectados e que exibe menos conexões com o restante da rede. Tais comunidades são mais comumente derivadas por algoritmos não supervisionados e podem ter muitos atributos diferentes. Além disso, a definição específica de uma comunidade difere entre os algoritmos.

O algoritmo de Girvan-Newman (GIRVAN; NEWMAN, 2002) é um método de detecção de comunidades em redes complexas. Ele pertence à classe de algoritmos que utilizam o conceito de modularidade para identificar comunidades em uma rede. A modularidade é uma medida que quantifica a diferença entre a densidade de arestas dentro de um conjunto de nós e a densidade esperada de um grafo aleatório (NEWMAN, 2010). Sua fórmula é dada por:

$$Q = \frac{1}{2M} \sum_{i,j} \left(a_{ij} - \frac{k_i k_j}{2M} \right) \delta(X_i, X_j) \tag{3.44}$$

em que Q é a modularidade, M é o número total de arestas na rede, a_{ij} é o elemento da matriz de adjacência que representa a conexão entre os nós i e j, k_i e k_j são os graus dos nós i e j, respectivamente, $\delta(X_i, X_j)$ é uma função delta que é 1 se os nós i e j pertencem à mesma comunidade e 0, caso contrário.

O algoritmo de Girvan-Newman segue os passos de remoção iterativa das arestas que mais contribuem para a modularidade da rede. Isso é feito removendo as arestas que estão associadas à medida de centralidade intermediância (GIRVAN; NEWMAN, 2002). A remoção de arestas com alta intermediância desfaz gradualmente a estrutura da rede, revelando comunidades distintas. Esse processo é repetido até que a estrutura da comunidade desejada seja revelada.

4 DESENVOLVIMENTO

Para modelar uma rede complexa, é preciso propor uma representação abstrata em um grafo, cujos nós (vértices) identificam os elementos do sistema e em que o conjunto de conexões (arestas) representa a presença de uma relação ou interação entre esses elementos. A partir dos dados da SSP-SP, realizou-se uma subdivisão do Estado de São Paulo em milhares de microrregiões (células), que juntas formam uma grade ou *grid* (c.f. Figura 21). As células foram usadas como elementos do sistema (nós do grafo).

Posteriormente, procedeu-se à agregação das ocorrências de roubo de veículos cujas coordenadas de latitude e longitude estavam contidas em uma mesma célula. A partir dessa consolidação, foram construídas séries temporais que registravam a quantidade de eventos (ocorrências) em um determinado período de tempo para cada célula. Para demarcar relação, considerou-se traçar a similaridade entre as séries temporais, já que estas descrevem bem a dinâmica criminal de cada microrregião. Duas células são similares se têm séries temporais similares, e neste caso, estão conectadas (por uma aresta).

Este capítulo apresenta a metodologia subdividida em etapas, cada uma abordando diferentes aspectos do estudo. O software que desempenhou um papel central em todas as operações computacionais realizadas neste trabalho foi a linguagem de programação Python[1], que combina uma sintaxe concisa e clara com os recursos poderosos de sua biblioteca padrão, além de módulos e *frameworks* desenvolvidos por terceiros.

4.1 Extração dos dados

Embora o processo de extração dos dados seja uma parte integrante da metodologia, os resultados finais independem da maneira pela qual a base de dados é obtida. A extração foi realizada utilizando a biblioteca Selenium[2], que possibilita a automação de interações com páginas da web. Um *script* foi desenvolvido para interagir com os elementos do site da transparência da Secretaria de Segurança Pública do Estado de São Paulo, permitindo a seleção dos botões necessários para o download de cada tabela associada a roubo de veículo. No entanto, é importante ressaltar que o site apresenta problemas relacionados à falta de padronização no design, inconsistências no *back-end* e questões de servidor e hospedagem, o que dificulta a ação de qualquer algoritmo de mineração.

Adicionalmente, foi preciso lidar com a presença de arquivos corrompidos, cuja leitura não podia ser feita diretamente. Para solucionar esse problema, cada planilha corrompida foi aberta manualmente e teve seu conteúdo copiado para um novo documento em formato ".xlsx", seguro, garantindo a integridade dos dados ali presentes. Em seguida, todos as planilhas foram fundidas em um único arquivo ".csv", agregando todas as instâncias das planilhas originais e formando a base de dados do trabalho.

[1] https://www.python.org/
[2] https://www.selenium.dev/

4.2 Seleção dos dados

Uma vez que a extração foi realizada, deferiu-se os procedimentos para a seleção na base de dados. Foi necessário traçar um período de interesse, remover duplicidades, tratar valores ausentes, e selecionar os campos com informações relevantes, conforme documentado. No processo de leitura, manipulação e seleção dos dados, a biblioteca Pandas[3] desempenhou um papel essencial. A partir dela, foi possível instanciar as planilhas oriundas do portal da transparência como objetos do pandas (*DataFrame*), que têm todos os métodos necessários.

4.2.1 Seleção das instâncias

Inicialmente, optou-se por não utilizar dados dos anos iniciais da coleta, de 2003 a 2010, devido à não completa implementação do sistema RDO, que registra e fornece os boletins de ocorrência, em todas as delegacias do estado. Além disso, havia inúmeros valores ausentes e registros inconsistentes nas datas iniciais. Essas limitações tornaram prudente considerar dados somente a partir de 2011.

Entretanto, eventos marcantes no intervalo de 2011 a 2021, como a Copa do Mundo de Futebol em 2014, ocorrida no Brasil, e a pandemia de COVID-19, que provocou mudanças substanciais nas políticas públicas entre os períodos em que se sucederam, poderiam introduzir padrões incomuns na dinâmica criminal (GOMES *et al.*, 2023), impactando negativamente a análise. Devido a esses fatores, o ano de 2022 foi selecionado, por apresentar pouca influência de eventos externos, como os mencionados. Além disso, é o ano mais recente com dados completos disponíveis. Após juntar todos os meses de 2022, um total de 124.512 ocorrências, representando instâncias na base, foram reunidas.

Cada linha constante na base registra os dados de uma pessoa, natureza ou objeto relacionado, o que significa que um boletim possuindo a identificação de mais de uma pessoa, natureza ou objeto possui mais de uma instância associada. Como o objetivo foi analisar o número bruto de ocorrências, foi removida a multiplicidade das instâncias relacionadas a um mesmo boletim. A remoção se deu por meio dos campos: 'NOME_DELEGACIA', 'ANO_BO', 'NUM_BO', conforme orientação da SSP-SP. Após o processo, restaram 52.554 ocorrências de roubos de veículo distintas entre si.

Mesmo restrita ao ano de 2022, a base possuía valores ausentes. Em particular, ausências nos campos de latitude e longitude tornavam as instâncias inviáveis para utilização, pois a componente espacial seria fundamental para agregar as ocorrências em células, conforme foi melhor discutido na Seção 4.2.2, e a escolha foi por preteri-las da análise. Das variáveis de interesse, também havia valores ausentes para a hora da ocorrência, o que impossibilitaria a criação de séries temporais com defasagem menor do que um dia. A fim de permitir uma análise que contemplasse as ocorrências para defasagens na unidade de hora, esses dados também foram excluídos da base final (para mais informações, confira Apêndice B, Seção B.2.1). Após a remoção das instâncias com valores ausentes, sobraram 46.231 ocorrências.

Um tratamento adicional foi necessário para a preparação das séries temporais, visto que

[3] https://pandas.pydata.org/

as tabelas incluíam boletins de delitos que ocorreram em anos anteriores e foram registrados apenas em 2022. Considerar esses delitos introduziria um viés significativo, pois estenderia as séries até o período em que eles ocorreram sem, contudo, considerar os outros crimes que se sucederam dentro da janela temporal aberta e foram notificados no seu ano de circunscrição, que são a maior parte dos casos. Dessa forma, optou-se por considerar apenas os boletins referentes a ocorrências registradas no período de 31 de dezembro de 2021 a 31 de dezembro de 2022, levando em consideração o campo 'DATAOCORRENCIA'. Após isso, sobraram 45.916 instâncias, sendo esse o número final de ocorrências analisadas.

4.2.2 Seleção das colunas

Após a seleção dos dados, restava escolher os campos pertinentes à análise. Optou-se por incluir as colunas 'BAIRRO', que continha informações sobre o bairro ou distrito onde o crime ocorreu, 'CIDADE', com informações sobre a cidade do incidente, 'LATITUDE' e 'LONGITUDE' para precisas coordenadas geoespaciais, e 'DATAOCORRENCIA' e 'HORAOCORRENCIA' para informações temporais específicas de data e hora, respectivamente.

As demais colunas não possuíam informação relevante para construção e análise das séries de eventos de cada região, ou não remetiam a fatores geoespaciais ou temporais das ocorrências. Para facilitar a manipulação dos dados, foi criada uma coluna, denominada 'DATAHORA', que aglutinou as informações de 'DATAOCORRENCIA' e 'HORAOCORRENCIA', norteando a criação da séries de eventos.

4.3 Criação do *grid*

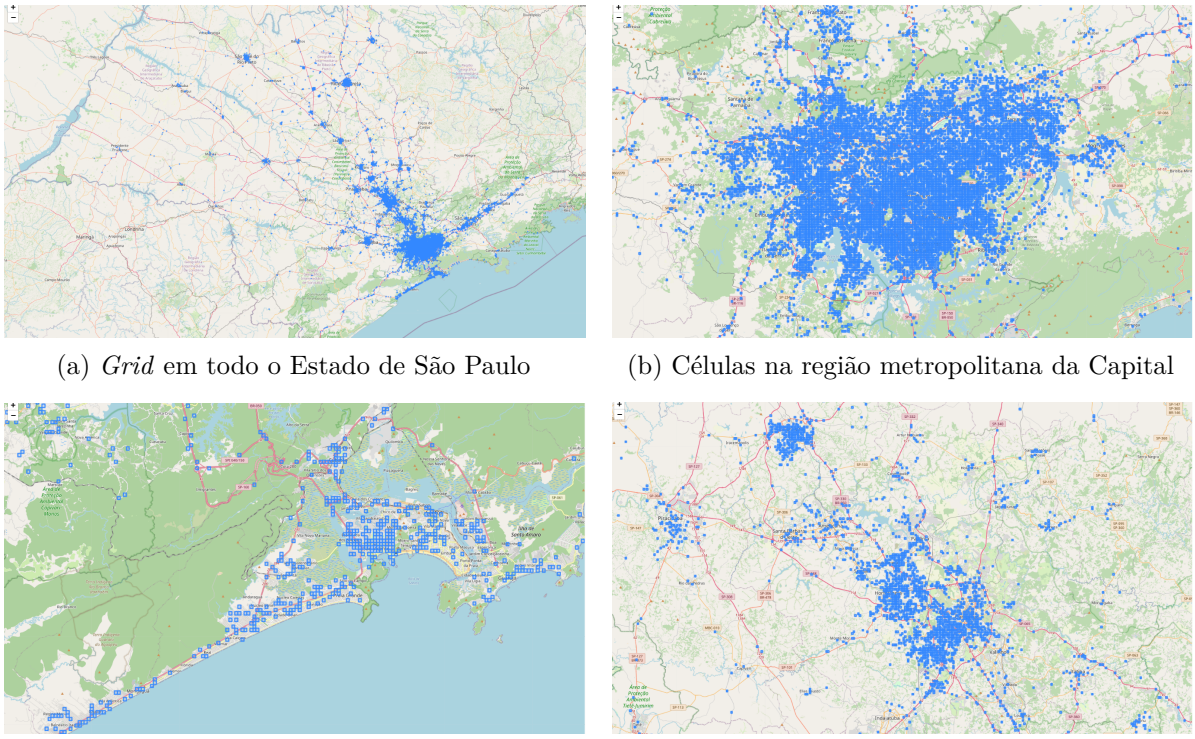
A partir das instâncias selecionadas, foi criado um *grid*, cujas células representam uma porção de área com alguma ocorrência registrada, para agregar eventos dentro uma mesma microrregião. Cada instância compreendida em uma região geográfica (definida pela latitude e longitude específicas da instância) gerou uma célula quadrada com proporções predefinidas e idênticas a todas as demais. Instâncias que residiam em regiões que já possuíam células geradas apenas foram agregadas a estas.

Ao todo, 17.421 células foram geradas (c.f Figuras 21 e 22). O *grid* foi esparso, à medida que não faria sentido definir regiões sem nenhum tipo de ocorrência com séries fossem vazias (sem eventos). Para a sua geração, as bibliotecas Geopandas[4] e Shapely[5] foram fundamentais. O Geopandas disponibiliza diversas ferramentas para manipulação de dados geoespaciais, enquanto o Shapely foi usado para facilitar a determinação da geometria das células a partir do objeto "Polygon".

As visualizações em mapa foram produzidas por meio da biblioteca Folium[6], que proporcionou uma integração simples e eficaz na criação de mapas interativos, facilitando a disposição e escolha das regiões a ser apresentadas, e permitindo uma interação dinâmica.[7]

[4] https://geopandas.org/
[5] https://shapely.readthedocs.io/
[6] https://python-visualization.github.io/folium/
[7] A maior parte das figuras elaboradas pelo autor que não apresentavam visualizações de mapas foi gerada pela biblioteca Matplotlib (https://matplotlib.org/)

Figura 21 – Visualização das células no mapa de São Paulo



(a) *Grid* em todo o Estado de São Paulo

(b) Células na região metropolitana da Capital

(c) Células na região da Baixada Santista

(d) Células na região de Campinas

Gerada pelo autor

As células no mapa bidimensional são representadas como quadrados, todos com o mesmo comprimento, cada um cobrindo uma área aproximada de cem mil metros quadrados. No entanto, é importante observar que, devido à natureza esférica da Terra, a área de formas geométricas de mesmas medidas em diferentes locais pode variar devido à curvatura característica. Essa variação é determinada pelas coordenadas de latitude e longitude associadas a cada local, influenciando as células geradas.

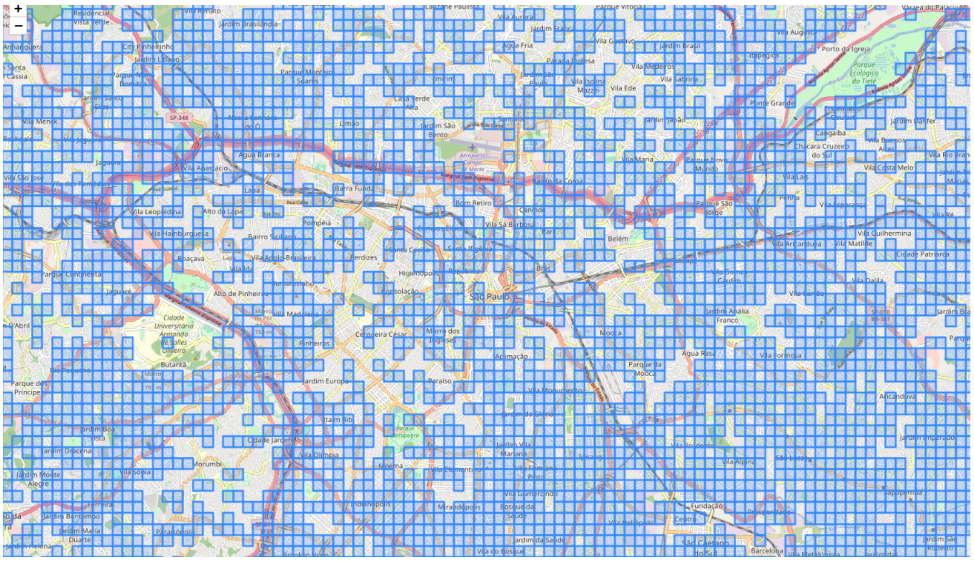
Tabela 3 – Estatísticas das áreas entre as células definidas no *grid*

Estatística descritiva	**Área (km^2)**
Média	0.1019
Desvio padrão	0.0005
Menor área	0.1008
Primeiro quartil	0.1017
Segundo quartil	0.1018
Terceiro quartil	0.1019
Maior área	0.1050

Fonte: Elaborada pelo autor.

No entanto, conforme evidenciado pela Tabela 3, a diferença entre as áreas das células geradas no processo é ínfima, especialmente quando consideramos que cada célula delineia uma região de crimes. Essa pequena diferença se deve à diminuta variação de latitude e longitude, uma vez que as células são extremamente pequenas em dimensão e estão confinadas dentro do

Figura 22 – Dimensão das células em comparação ao centro de São Paulo



Gerada pelo autor

território do Estado de São Paulo. Dessa forma, é seguro afirmar que nenhuma forma de viés decorrente da disparidade nos tamanhos das regiões analisadas impactou o resultado final. Para mais informações acerca do tema, recomenda-se Rheinwalt *et al.* (2012)

4.4 Criação das séries

Uma vez que as células e suas respectivas ocorrências estiveram bem definidas, como cada ocorrência possui informações de data e hora, foi possível criar séries temporais cujas variáveis contavam a quantidade de ocorrências em um determinado passo de tempo para cada célula. A criação da séries, assim como todos os outros procedimentos que envolveram computação científica, foram assistidos pela biblioteca Numpy[8].

4.4.1 Conceitos para a criação das séries

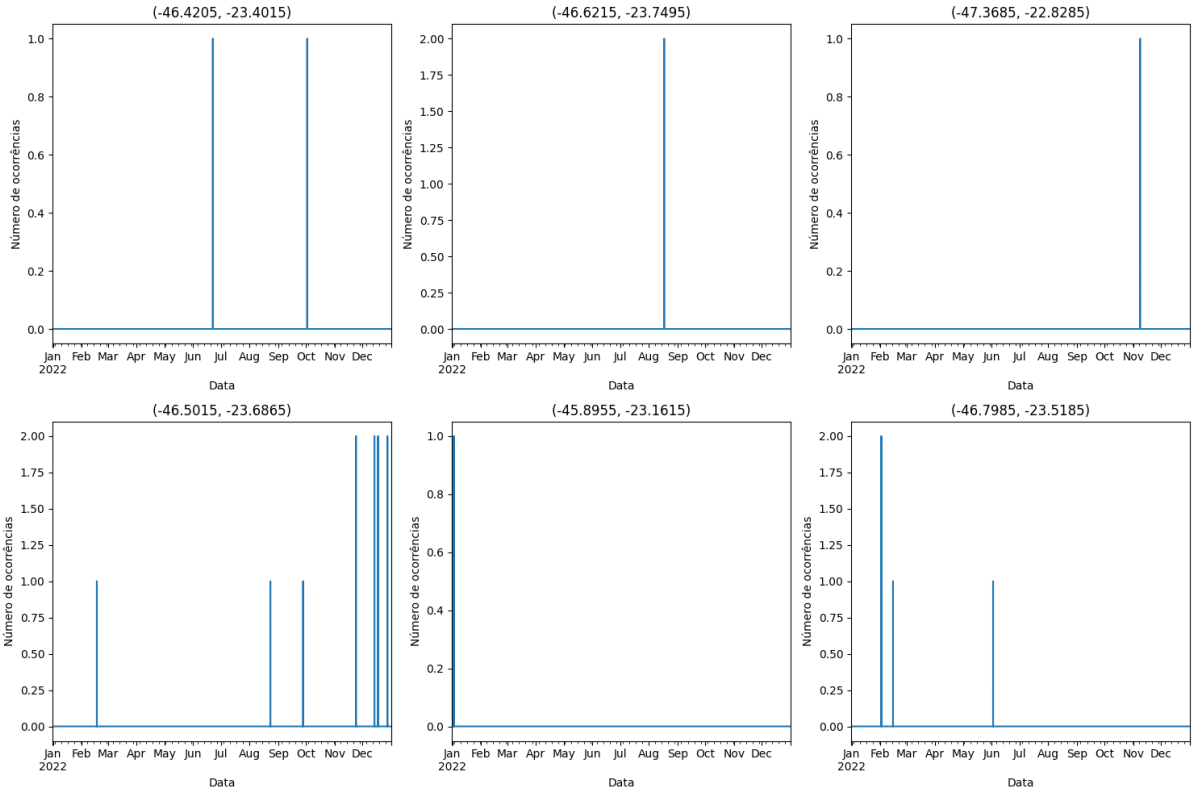
A partir das informações temporais dos eventos, construiu-se a série temporal de cada célula, num total de 17.421 séries temporais. As séries construídas possuíram uma variável, contando a quantidade de ocorrências de roubos de veículos registrada na região demarcada pela célula, e tiveram dimensão temporal, definida durante o período de 31 de dezembro de 2021 a 31 de dezembro de 2022. Para rotular as séries (e consequentemente suas respectivas células) o par ordenado '(longitude, latitude)' do centroide da célula foi usado.

Diferentemente da defasagem da Figura 23, mensal, para melhorar a visualização, a defasagem adotada para a criação das séries de eventos foi a horária (a série registra eventos durante o intervalo de uma hora, para cada hora do dia), para uma análise mais granular.

Com as séries temporais criadas, faltava apenas a definição de evento, para obter as séries de eventos. Como as células são suficientemente pequenas, e suas séries temporais são

[8] https://numpy.org/

Figura 23 – Exemplo com 6 séries temporais das células geradas



Gerada pelo autor

suficientemente esparsas[9], pôde-se definir que um determinado passo de tempo, isto é, um determinado horário do dia, possuiria um evento se uma ou mais ocorrências de roubo de veículo fossem registradas durante aquele período.

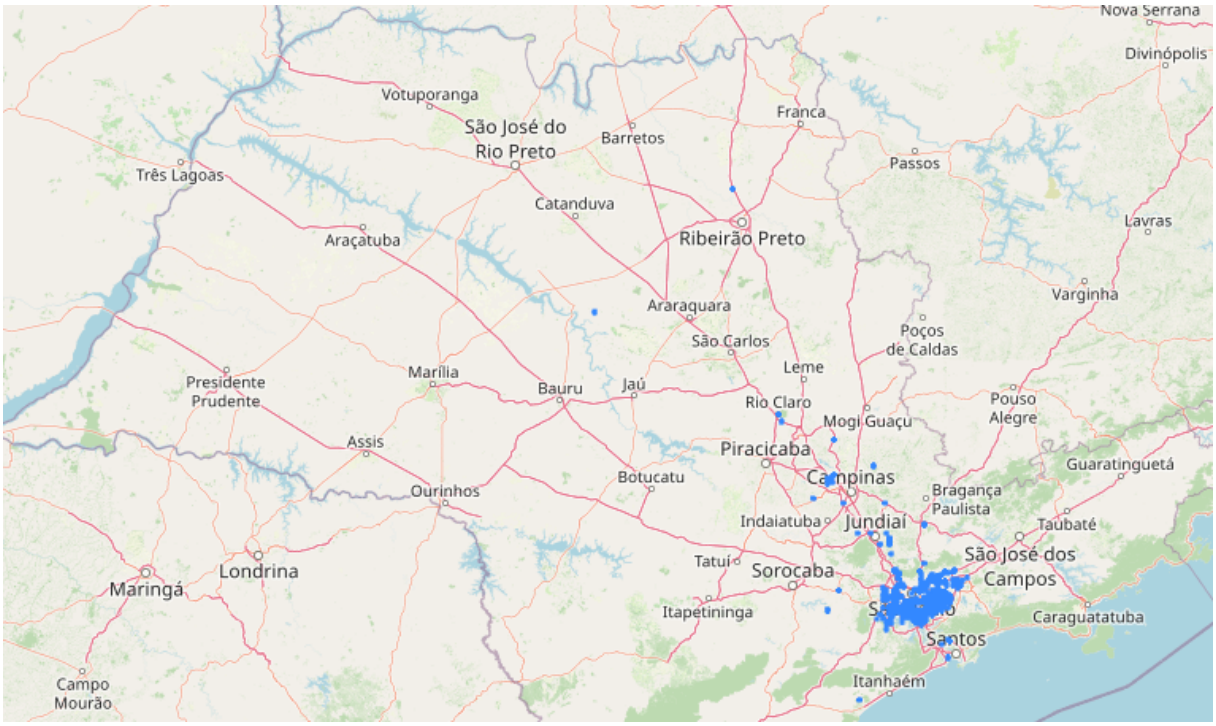
4.4.2 Filtros para as séries

Ao optar por séries temporais horárias, o próximo passo consistiu em filtrar as células, para direcionar os resultados a um escopo menos genérico. Foram escolhidas as células que apresentaram os 2% mais altos índices de eventos registrados, isto é, as células que possuíram as séries com a maior quantidade de eventos, acima de 98% da distribuição total.

Essa seleção permitiu concentrar os esforços de análise nas áreas geográficas que contribuíram de forma mais significativa para a quantidade total de ocorrências observadas em 2022, facilitando a identificação de padrões, tendências e características específicas que podem influenciar ou refletir aspectos importantes da atividade criminal no estado. Após essa filtragem, 433 células permaneceram do total original, apresentadas na Figura 24.

[9] Para séries horárias, com quase 9000 passos de tempo (aproximadamente a quantidade de horas em 366 dias), a célula com mais ocorrências registrou apenas 45 roubos de veículos. Para todas as células uma quantidade ínfima de eventos aconteceu em um mesmo passo de tempo.

Figura 24 – Células restantes após a filtragem



Gerada pelo autor

4.5 Aplicação de *Event Synchronization*

Com as instâncias selecionadas, as células criadas, e suas séries definidas e filtradas, para gerar as redes complexas, faltava apenas determinar a maneira pela qual as células se associariam. Pelos motivos discutidos nas Seções 3.2.2 e 3.3, decidiu-se que o critério de associação entre as células seria dado pela aplicação de *Event Synchronization*. Os cálculos de *Event Synchronization* e as simulações de Monte Carlo para análise de significância foram realizados com auxílio da biblioteca Pyunicorn[10], especializada em análise de séries eventos.

Para fins de comparação e devido à diferente natureza possível para os resultados, fez-se a aplicação do *Event Synchronization* simétrico e também do direcionado, com parâmetro $\tau_{max} = 252$. Assim sendo, uma matriz de adjacências simétrica e outra não simétrica foram geradas. Conforme descrito na Seção 3.3.1, para séries que apresentam taxas de eventos discrepantes, os valores para o ES não são comparáveis diretamente. A abordagem adotada foi a análise de significância, que baseia-se no método estatístico descrito na Seção 3.3.1. O objetivo primordial foi determinar se os padrões observados nas ocorrências de eventos são estatisticamente significativos, sugerindo uma possível sincronização entre eventos, ou se ocorrem de maneira aleatória. A análise de significância neste contexto envolveu o cálculo de valores-p associados às forças de sincronização Q_{ij} calculadas pelo ES.

Os valores-p representam a probabilidade de obter os Q_{ij} observados (ou mais extremos) apenas por acaso. O método emprega uma abordagem de Monte Carlo, onde séries de eventos simuladas são geradas com a mesma distribuição da série original de eventos, e as forças de

[10] https://pyunicorn.readthedocs.io/

sincronização são calculadas para essas séries simuladas a fim de estabelecer uma base para comparação. A avaliação é realizada comparando os Q_{ij} observados com aqueles obtidos a partir dos dados simulados. Os níveis de significância resultantes são representados como $1 -$ valores-p. Eles fornecem uma medida quantitativa da probabilidade de que os padrões de ES observados sejam atribuíveis apenas ao acaso.

Realizou-se tal análise de significância com 1000 simulações. Para o *Event Synchronization* simétrico, o resultado foi uma matriz simétrica, com valores variando entre 0 e 1. Em tal matriz, cada linha e cada coluna representam uma célula, e os valores indicam os níveis de significância associados à sincronização entre os eventos das células correspondentes. A medida ES de cada célula em relação a si mesma é definida como sendo igual a 0 (c.f Figura 25).

Da mesma forma, para o *Event Synchronization* direcionado, a análise de significância proporcionou uma matriz não simétrica, também com valores entre 0 e 1, cujos níveis de significância dependiam da direção em que se pretendia enxergar as relações (c.f Figura 26). Esses resultados constituem uma base quantitativa para avaliar a confiabilidade dos padrões observados de sincronização em relação ao acaso.

Figura 25 – Exemplo de resultado para análise de significância do caso simétrico. O valor a_{ij} representa a significância da sincronização entre as células i e j

	Célula 'A'	Célula 'B'	Célula 'C' ...	
Célula 'A'	0	0.13	0.3	...
Célula 'B'	0.13	0	0.65	...
Célula 'C'	0.3	0.65	0	...
⋮	⋮	⋮	⋮	⋱

Figura ilustrativa gerada pelo autor

Figura 26 – Exemplo de resultado para análise de significância do caso direcionado. O valor a_{ij} representa a significância da sincronização entre as células i e j

	Célula 'A'	Célula 'B'	Célula 'C' ...	
Célula 'A'	0	0.41	0.89	...
Célula 'B'	0.53	0	0.91	...
Célula 'C'	0.23	0.11	0	...
⋮	⋮	⋮	⋮	⋱

Figura ilustrativa gerada pelo autor

4.6 Criação das redes

Uma vez que os resultados da análise de significância foram obtidos, pôde-se determinar um limiar de nível de significância que represente ou não uma sincronia entre duas regiões, isto é, uma valor para o qual considera-se que o resultado obtido considerando as 1000 simulações possa ser tido como uma evidência plausível de sincronia entre os eventos das duas células. Para

operações envolvendo redes, a biblioteca NetworkX[11] foi empregada, oferecendo uma variedade de algoritmos e ferramentas para análise e visualização de redes complexas. Toda a manipulação necessária e todas as medidas locais e globais expostas no Capítulo 5.1 foram obtidas a partir dela, assim como as comunidades.

4.6.1 Matrizes de adjacências

A partir das matrizes resultantes da análise de significância, criou-se as matrizes de adjacência que dariam origem às redes que se pretendia analisar. Para tanto, considerou-se que um nível de significância maior ou igual a 0,95 (valor-p menor ou igual a 0,05) representaria a existência de uma aresta entre dois nós. Isto é, para valores entre células maiores ou iguais 0,95, colocou-se 1 na matriz de adjacências, e 0, caso contrário, tanto para os resultados do ES simétrico quanto para os resultados do caso direcionado.

Uma matriz de adjacências define totalmente a topologia de uma rede complexa, contendo as informações dos nós e arestas em questão (c.f. Seção 3.4). Os procedimentos anteriores foram suficientes, portanto, para gerar as redes que foram discutidas nos resultados do Capítulo 5. Seus nós representam as células, que são as regiões com mais ocorrências de roubo de veículos do Estado de São Paulo, e suas arestas representam a existência sincronização entre as séries de eventos das células conectadas.

[11] https://networkx.github.io/

5 RESULTADOS E DISCUSSÃO

Nas Tabelas 4 e 5, encontram-se os resultados para os cálculos relacionados ao agrupamento temporal das séries de eventos das células remanescentes, após a filtragem.

Tabela 4 – Estatísticas do coeficiente de emparelhamento para as séries remanescentes

Estatística	Valor
Média	0.005
Desvio Padrão	0.025
Mínimo	0.0
25º Percentil	0.0
50º Percentil (Mediana)	0.0
75º Percentil	0.0
90º Percentil	0.0
95º Percentil	0.030
99º Percentil	0.143
Máximo	0.2

Tabela 5 – Estatísticas da taxa de agrupamento para as séries remanescentes

Estatística	Valor
Média	0.01
Desvio Padrão	0.044
Mínimo	0.0
25º Percentil	0.0
50º Percentil (Mediana)	0.0
75º Percentil	0.0
90º Percentil	0.0
95º Percentil	0.057
99º Percentil	0.25
Máximo	0.363

Fonte: Elaboradas pelo autor

Como é possível observar, ambas as medidas de agrupamento temporal são bastante baixas para as séries tratadas como um todo. De fato, o agrupamento temporal afeta muito pouco menos de 10% do total delas. Portanto, não se espera que ocorra nenhum viés nos resultados devido a esse fator.

Antes de explorar os resultados, é relevante destacar que todas as figuras apresentadas nas seções seguintes são derivadas de visualizações interativas disponíveis de forma detalhada em https://colab.research.google.com/drive/1_LSg65faUbbET24Nqsb6FGP5g3p_hMzg?authuser=1. Para fazer o download dessas figuras, basta acessar https://drive.google.com/drive/folders/1dCTPEEufIQDhh7WOceJqnaLRx1qtzEwG?usp=drive_link.

5.1 Resultados

5.1.1 Rede simétrica

Iniciando a análise dos resultados, apresentam-se as medidas expostas na Seção 3.4 para a rede simétrica. O estudo detalhado da rede, que valeu-se de medidas globais, locais e da aplicação de algoritmos para formação de comunidades (Girvan-Newman), revelou informações valiosas sobre sua estrutura e dinâmica, como destacado na Tabela 6.

Tabela 6 – Medidas globais para a rede simétrica

Medida Global	**Valor**
Densidade de Conexões	0,015
Comprimento médio do caminho mais curto	1,744
Raio da maior componente conexa	4
Diâmetro da maior componente conexa	7
Transitividade	0,128
Assortatividade	0,012

Fonte: Elaborada pelo autor.

A densidade de conexões, notavelmente baixa com um valor de 0,015, indica que apenas um pouco mais de 1% do total de conexões possíveis na rede foi estabelecido. Esse número propõe que os nós não se relacionam com facilidade. Isto é, as diferentes regiões geográficas não apresentam semelhança significativa, grosso modo, o que ressalta as particularidades das ocorrências criminais.

Ao analisar o comprimento médio do caminho mais curto, observa-se um valor de 1,744. Essa medida sugere uma eficiência notável no alcance entre os nós da rede. No entanto, é importante ressaltar que o conceito de ”caminho mais curto” só se aplica aos nós dentro da mesma componente conexa. Com apenas três componentes conexas, das quais uma é significativamente maior com 362 nós, e as outras duas possuem apenas 2 nós cada, cerca de 67 nós não estão conectados a nenhum outro.

A caracterização topológica da maior componente conexa revelou um raio de 4 arestas e um diâmetro de 7 arestas. Esses valores propõem a extensão e alcance dessa componente específica, pois representam o tamanho médio dos caminhos mais curtos e a distância máxima entre os nós, respectivamente. Como uma aresta significa sincronização entre as séries de duas regiões, tal resultado indica que células que sincronizam com outras numa mesma estrutura podem estar a até 7 graus de separação.

Com relação à transitividade, que mede a propensão de conexões entre os vizinhos de um nó, o valor de 0,128 sugere uma tendência moderada para a frequência com que as vizinhas de uma célula estão conectadas entre si. A sincronia proposta pelo *Event Synchronization* não é transitiva, no sentido que a conexão dos vizinhos de um nó entre si não é necessária (como se espera de sincronizações no sentido físico).

Por outro lado, a assortatividade, que mede a preferência dos nós por conexões com pares de grau semelhante, retornou um valor de 0,012, muito baixo, indicando que não há relação entre a quantidade de sincronias da célula com a de suas vizinhas. No escopo do trabalho, cabe dizer que em média, as células têm padrão criminal que sincroniza igualmente com tendências mais e menos comuns, sem qualquer predileção

A Tabela 7 apresenta as células com maiores medidas locais observadas. Elas são muito relevantes para traçar os nós tidos como ”centros” ou ”eixos”, que desempenham um papel fundamental na conectividade e na eficiência da comunicação dentro da rede.

Vale a pena destacar que como as células representam uma porção geográfica, fez-se

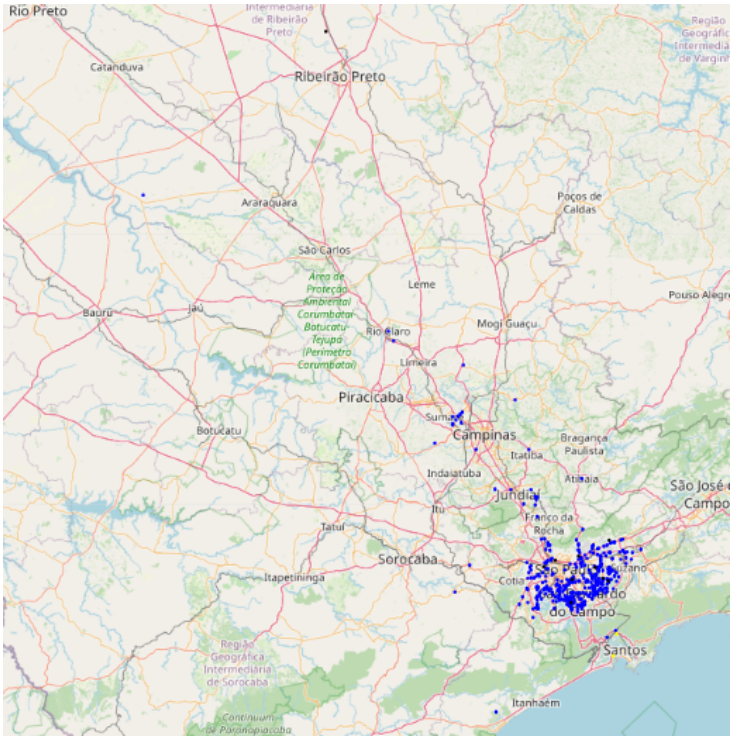
Tabela 7 – Células com maiores medidas locais

Medida Local	**Célula (longitude, latitude)**
Grau	(-46,7535, -23,6505)
Intermediância	(-46,5405, -23,5935)
Coeficiente de agrupamento local	(-46,5495, -23,7105), (-46,4775, -23,6685), (-46,3185, -23,4675), (-46,2885, -23,4705), (-46,8165, -23,5335), (-46,4835, -23,5755)

Fonte: Elaborada pelo autor.

interessante apresentá-las a partir de mapas. Contudo, para avaliar os resultados, a visualização em mapas pode ser prejudicada pela dimensão das células, pequena demais em comparação à extensão do território que as compreende no *grid*, conforme ilustra a Figura 27.

Figura 27 – Comunidades na rede considerando células

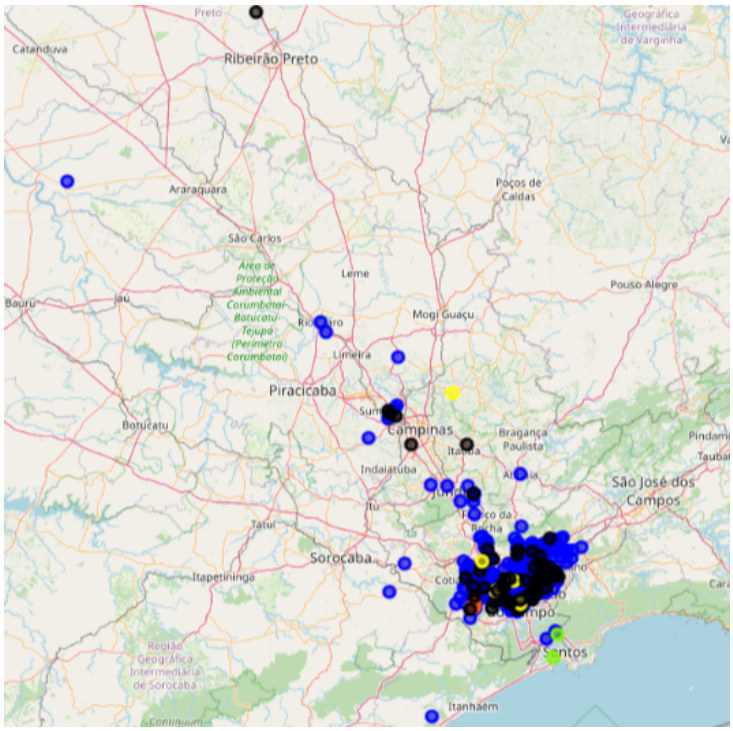


Gerada pelo autor

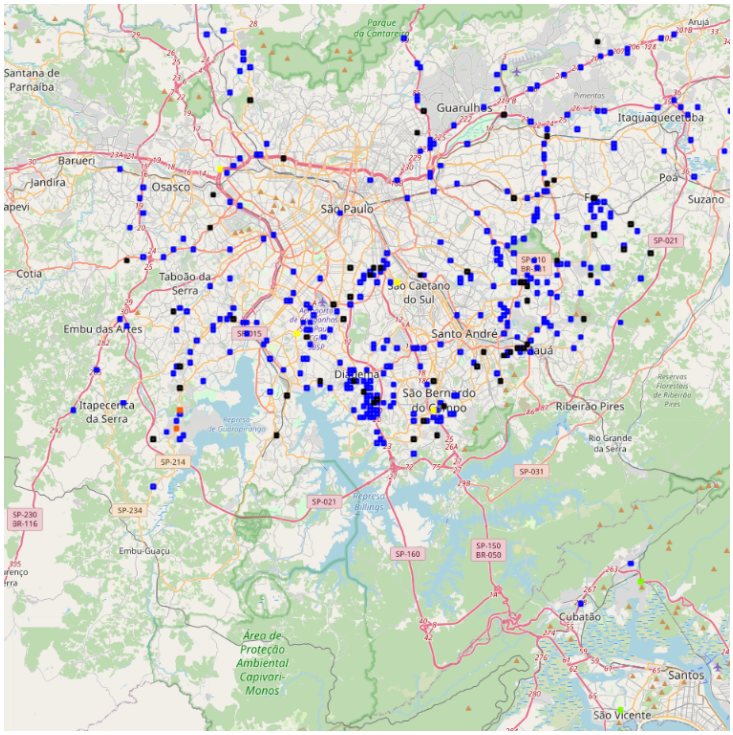
A solução para isso foi elaborar mapas com pontos simbólicos, mais largos que as células, porém de mesma localização que as mesmas, para representá-las (c.f. Figura 28a). Nos contextos em que foram avaliadas regiões específicas do Estado de São Paulo, notadamente a capital, fez-se uso das células em seu tamanho real, conforme Figura 28b.

As cores das células representam a comunidade em que elas estão inseridas, com exceção do preto, que representa uma célula que está em uma comunidade com um único elemento (ela própria). Para o contexto do presente trabalho, uma célula numa comunidade de um elemento, na verdade, não possui comunidade.

Figura 28 – Comunidades na rede simétrica



(a) Em Pontos simbólicos



(b) Nas células da capital

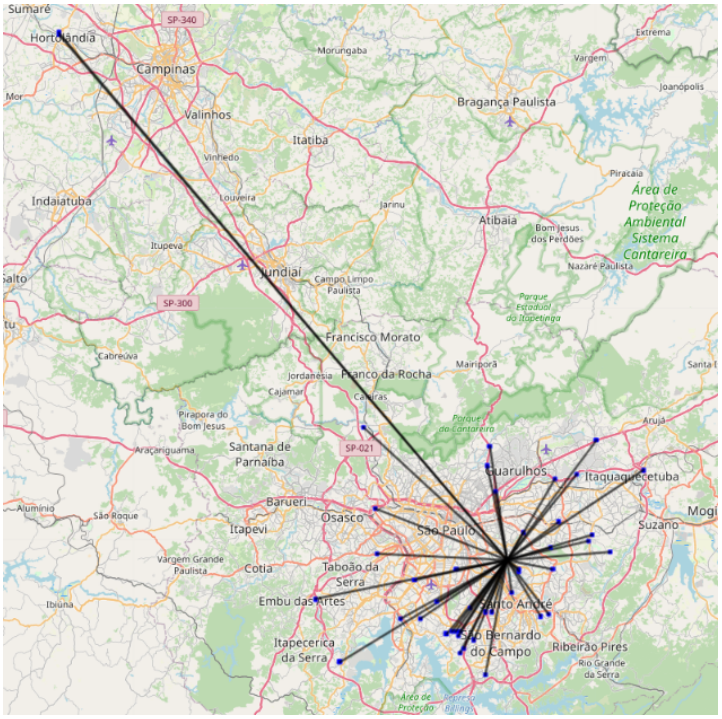
Gerada pelo autor

O algoritmo de Girvan-Newman opera removendo arestas da rede com base em suas contribuições para a intermediância. No contexto do estudo, as arestas removidas tendem a ser aquelas que conectam duas células cujos vizinhos são distintos. Quatro comunidades foram

formadas: Uma em azul, que compreende a maior parte das células, 357; uma em amarelo com um total de 5 células, sendo 4 na região metropolitana da capital e 1 em Pedreira, macrorregião de campinas; uma em laranja, com 2 células, ambas localizadas no Jardim Ângela, na Capital; e uma em verde, com duas células na região da baixada, precisamente nas cidades de Cubatão e São Vicente. As demais 67 células não possuem comunidades: estas são, na verdade, exatamente as 67 células que não estão em nenhuma componente conexa e, portanto, não se ligam a nenhuma outra.

Na Figura 29, desenvolveu-se uma visualização para a célula (-46,5405, -23,5935), que aparece na Tabela 6, com suas arestas e células ligantes. A célula de interesse aparece em vermelho, apesar da cor ser difícil discernir devido aos demais elementos na visualização. As células ligantes estão em azul.

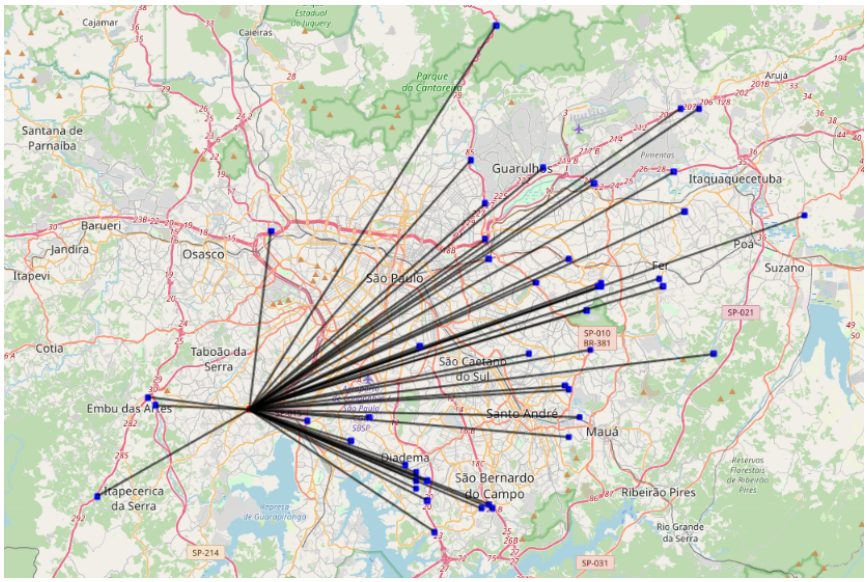
Figura 29 – Arestas e nós ligados ao nó centrado em (-46,5405, -23,5935)



Gerada pelo autor

Esta célula, com maior centralidade de intermediância na rede, situa-se na cidade de São Paulo, no distrito São Lucas. Ela liga-se a células nas zonas sudoeste, sudeste e leste, além de distritos no município de São Paulo. Na região metropolitana, além da capital, há arestas com outras células em Embu das Artes, Itapecerica da Serra, Diadema, São Bernardo do Campo, Arandu, Mauá, Guarulhos e Itaquaquecetuba. Duas únicas células aparecem fora da região metropolitana, localizadas em Hortolândia, vizinhas geograficamente. É interessante notar, desde já, que a região metropolitana da capital, como era de se esperar (Bulletin d'information sur la criminalité et l'organisation policière, 2000), acumula a maior parte das células presentes no estudo e as mais centrais também.

Figura 30 – Arestas e nós ligados ao nó centrado em (-46,7535, -23,6505)



Gerada pelo autor

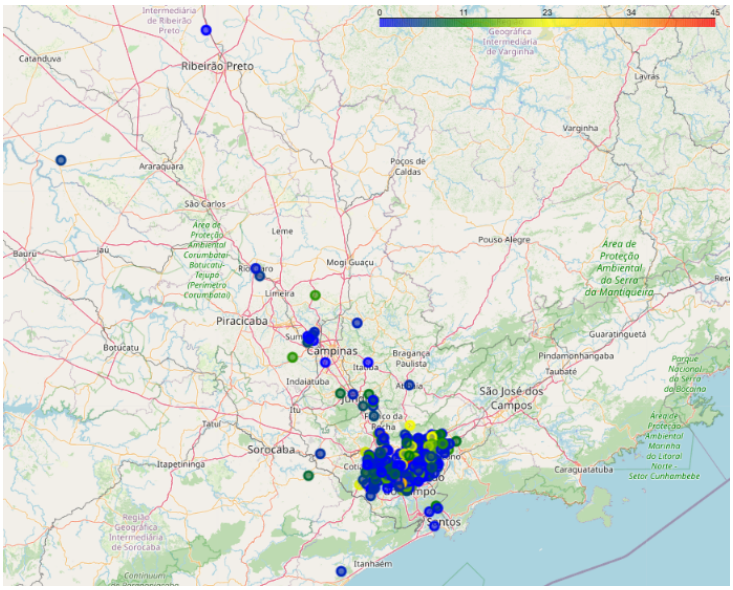
A partir da Figura 30, pôde-se observar que a célula de maior centralidade de grau, (-46.7535, -23.6505), situa-se na cidade de São Paulo, no distrito Jardim São Luis. Ela liga-se exclusivamente a células na região metropolitana da capital, nas regiões sudoeste, sudeste e leste, além de distritos no município de São Paulo. As cidades associadas às células que aparecem Embu das Artes, Itapecerica da Serra, Diadema, São Bernardo do Campo, Arandu, Mauá, Guarulhos e Suzano. É possível notar algumas localidades em rodovias que cortam o estado, fora de regiões urbanas.

Na Figura 31, uma visualização para os graus na rede foi desenvolvida a partir do mapa com as posições de cada célula, em que as cores não representam mais as comunidades, mas sim seu valor para a medida no contexto da escala de valores observados, que pode ser conferida no canto superior direito da imagem. Cores frias representam valores mais baixos e cores quentes representam valores mais altos.

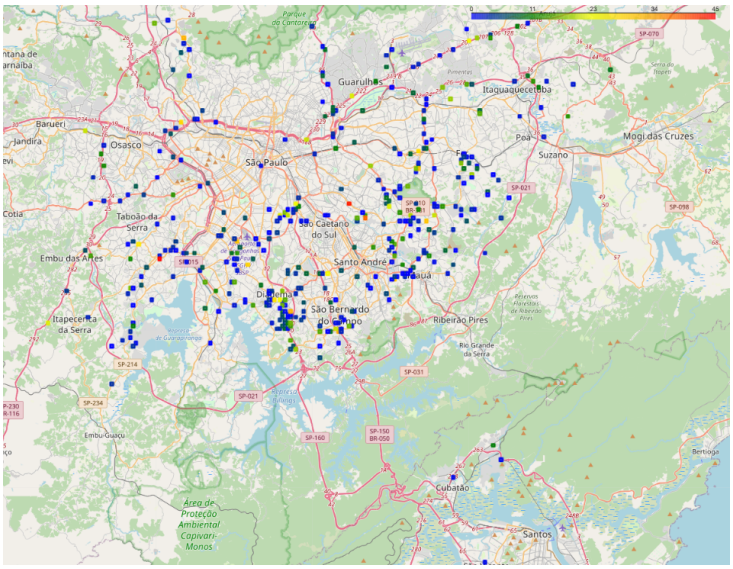
Os graus, ilustrados na Figura 31, têm uma distribuição mais ou menos uniforme, com muitas células de grau baixo ou nulo (o que indica nenhuma sincronização), algumas poucas células com grau intermediário e menos de dez células com grau igual ou superior a 30. Células sem arestas possuem séries de eventos que não foram suficientemente similares, dentro da metodologia do trabalho, isto é, não sincronizaram com nenhuma outra, mostrando um padrão incomum ao contexto. Por outro lado, a célula mais conectada, (-46,7535, -23,6505), possui 45 arestas, e portanto sua série de eventos é síncrona a outras 45 células presentes no *grid*.

A intermediância teve uma distribuição ainda mais uniforme, conforme Figura 32, com a maior parte dos valores muito menor do que o maior valor registrado, atingido pela célula (-46,5405, -23,5935). Justamente as duas células destacadas como eixos aqui assumem valores expressivos pro contexto, com poucas outras células com um valor alto. A região metropolitana

Figura 31 – Distribuição dos graus na rede simétrica



(a) Em todo o estado



(b) Nas células da capital

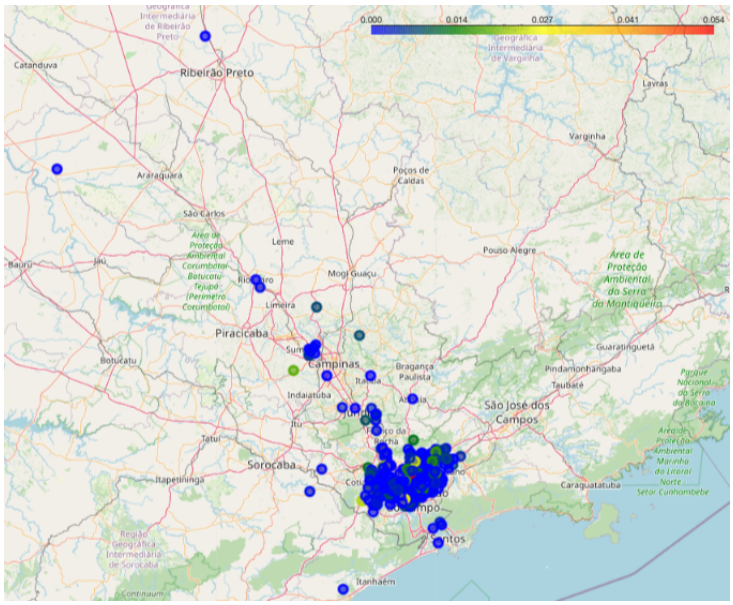
Gerada pelo autor

da capital contém a maior parte das células que atingiram os maiores valores.

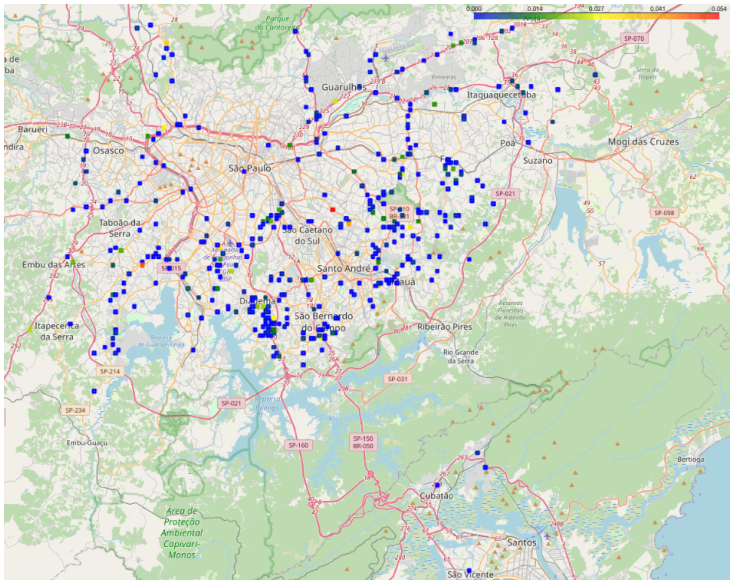
Como o coeficiente de agrupamento local indica o quão conectadas entre si estão as conexões de uma determinada célula, ele é uma medida local que fala sobre a estrutura da rede também. Nesse sentido, ter muitas conexões diminui o valor esperado para os eixos, a menos de redes densamente conectadas, que a tabela 6 mostra não ser o caso.

O coeficiente de agrupamento local revelou, na Figura 33, para a maior parte dos nós, uma porção pequena de suas conexões estão ligadas entre si. Contudo, alguns nós da capital atingiram valores razoáveis dessa medida. As células centradas em (-46,5495, -23,7105),(-46,4775, -23,6685), (-46,3185, -23,4675), (-46,2885, -23,4705), (-46,8165, -23,5335) e (-46,4835, -23,5755)

Figura 32 – Distribuição da intermediância na rede simétrica



(a) Em todo o estado



(b) Nas células da capital

Gerada pelo autor

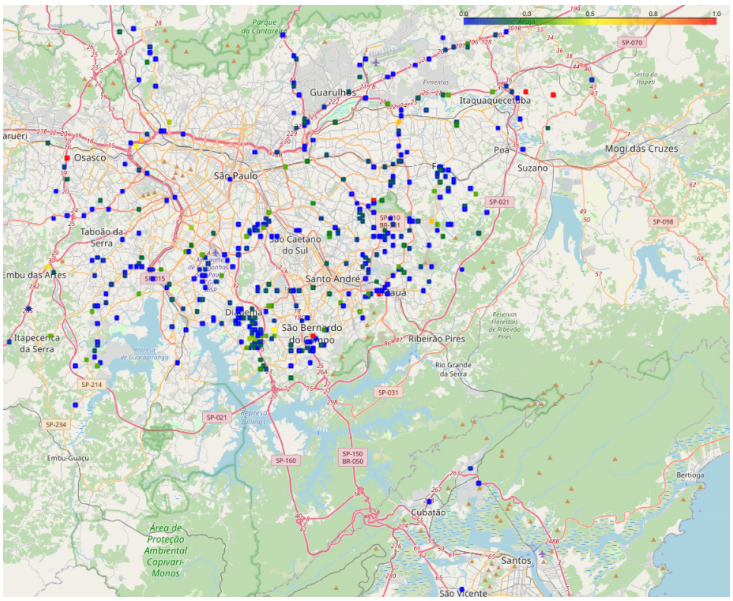
todas possuíram coeficiente de agrupamento local igual a 1. Este é um indicativo de que as células e suas conexões possuem, de fato, um padrão comum, já que o ES foi capaz de perceber a sincronização delas com seus ligantes e dos ligantes entre si.

5.1.2 Rede direcionada

Para a rede direcionada, um estudo similar foi desenvolvido, com as devidas alterações necessárias para interpretar a direcionalidade das ligações, além de novas métricas emergentes no contexto. A Tabela 8 as medidas globais do caso direcionado.

A densidade de links, representada pelo valor de 0,017, revela que cerca de 1,7% das

Figura 33 – Distribuição do agrupamento local nas células da capital para rede simétrica



Gerada pelo autor

Tabela 8 – Medidas globais para rede direcionada

Medida Global	**Valor**
Densidade das conexões	0,017
Comprimento médio do caminho mais curto	3,165
Raio da maior componente conexa	4
Diâmetro da maior componente conexa	7
Transitividade	0,062
Assortatividade	0,022
Reciprocidade	0,005

Fonte: Elaborada pelo autor.

conexões potenciais na rede foram efetivamente estabelecidas. Esse pequeno aumento percentual em relação ao caso simétrico, contudo, representa mais do que pode parecer, na comparação entre as duas redes. Como as arestas agora possuem direção, sua quantidade máxima possível, presente no cálculo da densidade, dobrou. Se o percentual aumentou, a quantidade nominal de arestas mais que dobrou.

O comprimento médio do caminho mais curto retornou um valor de 3,165. Esse valor propõe uma distância média relativamente curta entre os nós, revelando eficiência na comunicação e alcance na rede. O resultado vale para a única componente fortemente conexa registrada, que possui 319 células, com raio de 4 arestas e diâmetro de 7 arestas. Todas as outras células não estavam em qualquer componente fortemente conexa. Também há apenas uma componente fracamente conexa, com 423 células, com basicamente as mesmas células da componente fortemente conexa formando-a.

No que se refere à transitividade, o valor calculado de 0,062 denota uma diminuição em

comparação com a rede simétrica. Isso se deve, possivelmente, à direcionalidade das arestas, que agora representam um caminho, por exemplo, que vai da célula "A", passa pela "B", chega a "C" e retorna à célula "A". Ou seja, agora é preciso que as direções do triângulo formado saíam e cheguem à mesma célula. Já a assortatividade, apresentou um valor de 0,022.

A reciprocidade calculada foi de 0,005, um valor consideravelmente baixo. Isso ressalta uma proporção reduzida de conexões bidirecionais na rede, indicando que, em geral, a tendência é que eventos em uma série sejam sucedidos por eventos na outra, mas não o contrário. Em outras palavras, os padrões de sincronia são predominantemente unidirecionais: em mais de 99% das sincronias observadas, eventos sincrônicos ocorreram, em sua maioria, inicialmente em uma célula e, posteriormente, na outra, sem ocorrência significativa da situação oposta.

Tabela 9 – Células com maiores medidas locais

Medida Local	**Célula (longitude, latitude)**
Grau de entrada	(-46,5525, -23,7255)
Grau de saída	(-46,5405, -23,5935)
Intermediância	(-46,5405, -23,5935)
Coeficiente de agrupamento local de entrada	(-46,5525, -23,7255)
Coeficiente de agrupamento local de saída	(-46,5405, -23,5935)

Fonte: Elaborada pelo autor.

Para a rede direcionada, o grau e o coeficiente de agrupamento acabam se bifurcando em duas medidas, uma para entrada e outra para saída, devido à direcionalidade das arestas. É interessante notar que mesmo com uma variedade maior de medidas de centralidade avaliadas, apenas duas células distintas aparecem na Tabela 9. A rede, portanto, possui pontos centrais bastante manifestos.

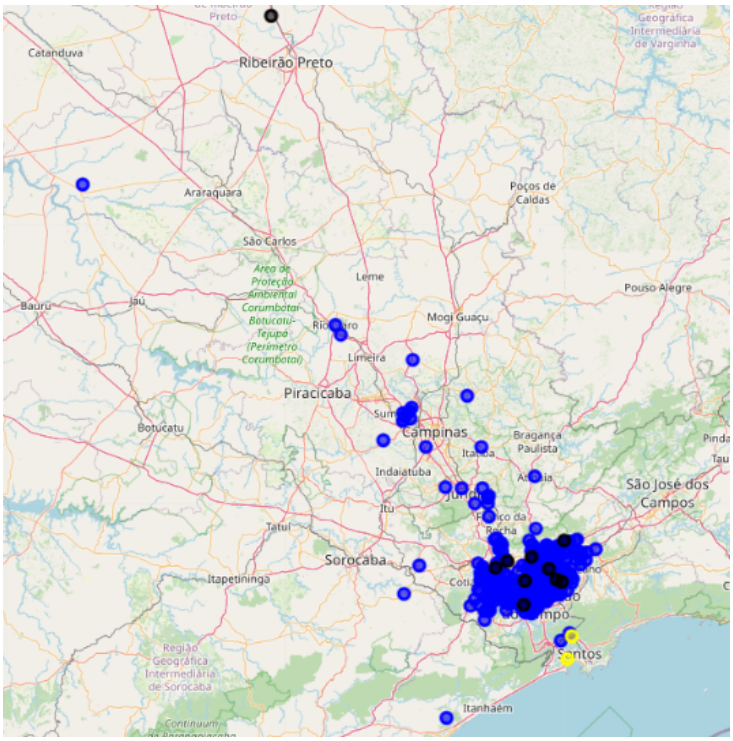
Analisando as comunidades a partir da Figura 34, é possível notar que a rede direcionada tem um padrão muito mais homogêneo que a simétrica, com a maior parte das células pertencendo à mesma comunidade, em azul. De fato, há apenas duas comunidades, a azul, que compreende a maior parte das células, com 421 elementos, e a amarela, com os exatos 2 elementos na baixada que apareceram também na rede simétrica. Apenas 10 células permaneceram sem nenhuma comunidade, o que mostra que de fato a coesão da rede aumentou bastante.

Na representação da rede direcionada, houve uma reestruturação das visualizações de nós e arestas para acomodar a presença de arestas de entrada e saída. Essa diferenciação foi necessária para aprimorar a clareza na interpretação dos elementos visuais. As arestas azuis são de saída para a célula em análise, enquanto as arestas de entrada estão destacadas em vermelho, acompanhadas por setas com a direção do fluxo. As Figuras 35 e 36 apresentam as células com maiores centralidades na rede direcionada.

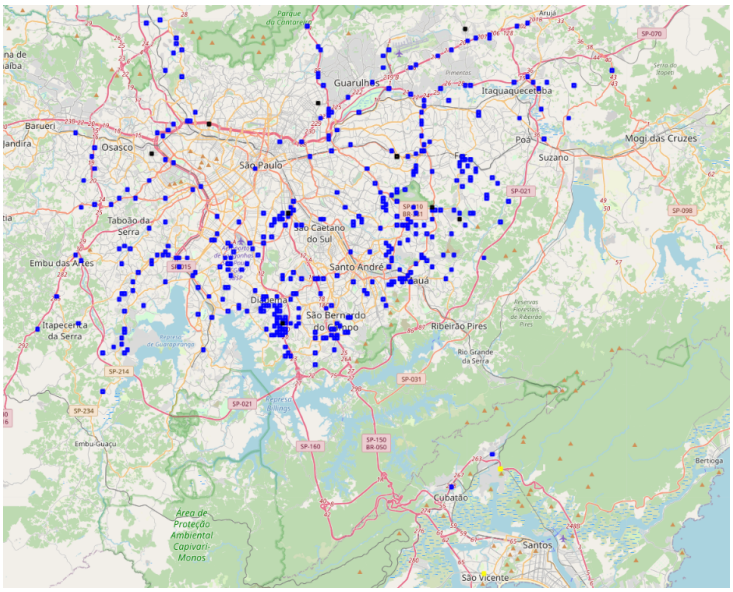
A célula centrada em (-46.5525, -23.7255) teve maior grau de entrada e maior coeficiente de agrupamento local de entrada na rede. Ela está localizada em São Bernardo do Campo, entre os bairros Demarchi e Ferrazópolis.

A célula (-46,5405, -23,5935), localizada em São Paulo, distrito São Lucas, que já havia aparecido na rede simétrica, obteve o maior valor para todas as demais medidas da rede. A

Figura 34 – Comunidades na rede direcionada



(a) Em todo o estado



(b) Nas células da capital

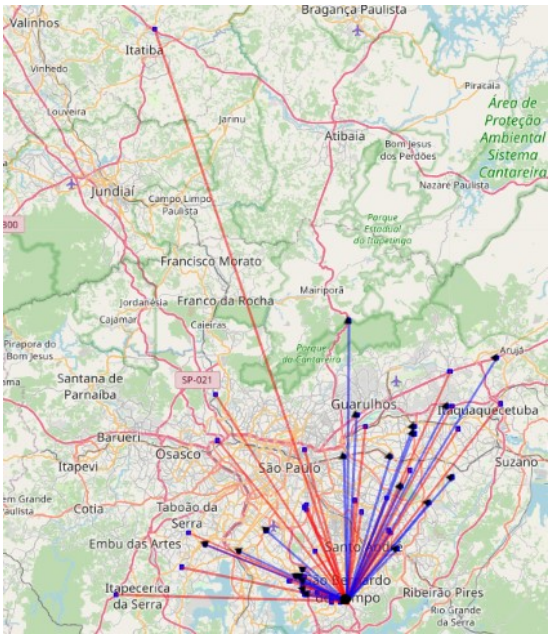
Gerada pelo autor

quantidade de elementos no mapa aumentou consideravelmente, o que dificulta a análise.

A Figura 37 ilustra todos os tipos de grau para células, com enfoque na região da capital, seguindo a mesma ideia de cores frias e quentes denotando o valor para a medida.

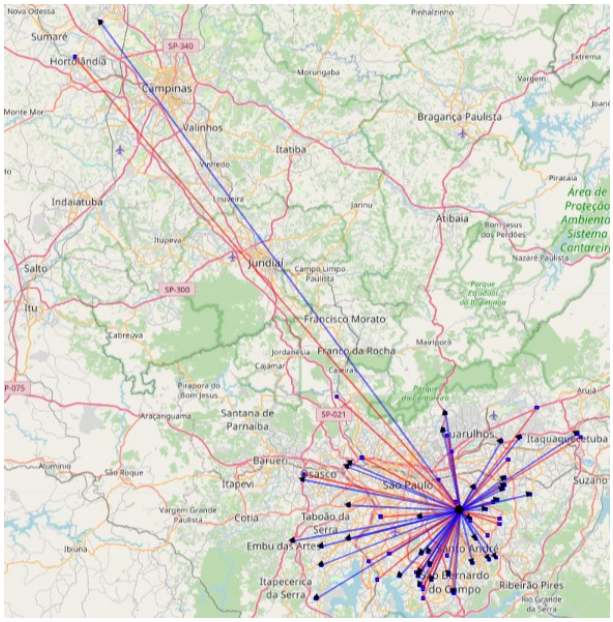
Os graus de entrada indicam a quantidade de arestas que chegam à célula. No contexto do trabalho, isso pode revelar a tendência de eventos acontecerem em outras lugares, antes do

Figura 35 – Arestas e nós ligados ao nó centrado em (-46.5525, -23.7255)



Gerada pelo autor

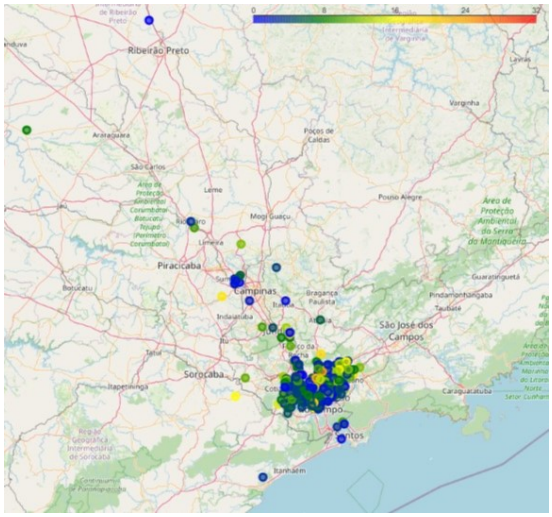
Figura 36 – Arestas e nós ligados ao nó centrado em (-46.5405, -23.5935)



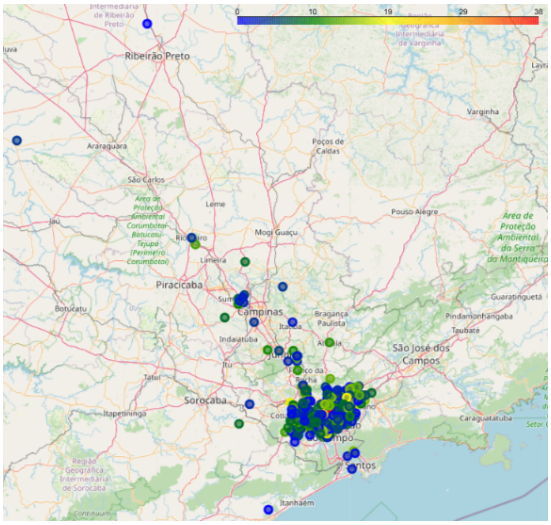
Gerada pelo autor

lugar em questão. Na distribuição dos graus de entrada, foi possível notar que muitas células possuem um grau intermediário dentro da faixa de valores possível. Isso evidencia, novamente, o caráter mais homogêneo da rede direcionada. Para os graus de saída, o raciocínio é oposto: as células com grau de saída alto tendem a ter eventos em si que são seguidos por eventos em outras

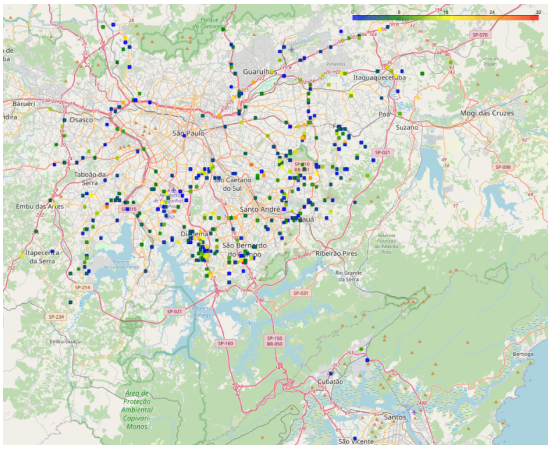
Figura 37 – Distribuição dos graus na rede direcionada



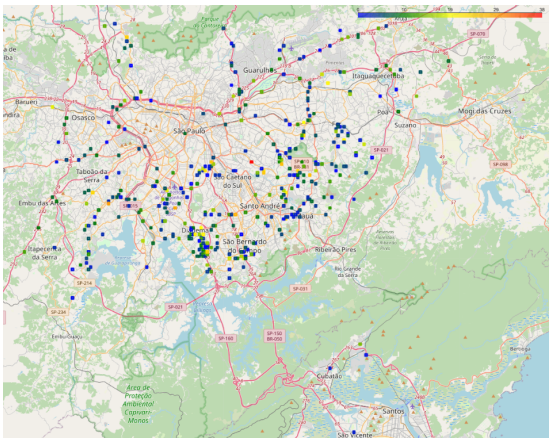
(a) Graus de entrada em todo o estado



(b) Graus de saída em todo o estado



(c) Graus de entrada nas células da capital



(d) Graus de saída nas células da capital

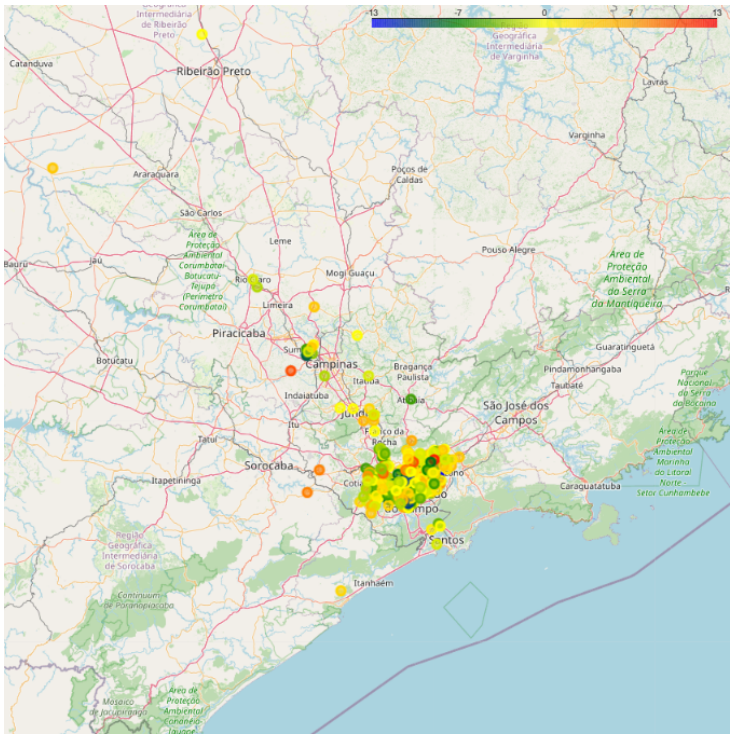
Gerada pelo autor

células. Na distribuição encontrada, boa parte dos locais permaneceram em faixas intermediárias, mas o número de graus baixos diminuiu, comparativamente.

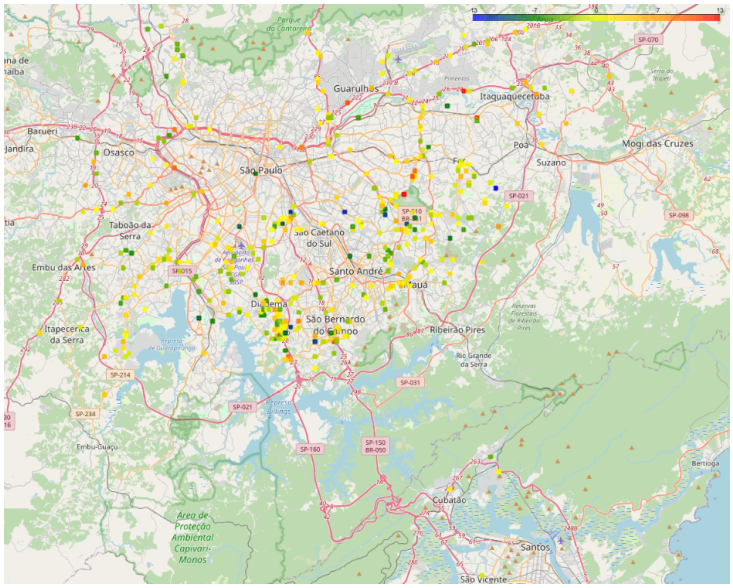
Na distribuição presente na Figura 38, foi possível notar que a maior parte dos locais apresenta valores baixos para a divergência. Uma divergência alta indica que há muitas arestas entrando, e relativamente poucas saindo. Assim sendo, com a divergência, é possível determinar as células que tendem anteceder ou suceder eventos, grosso modo, em outros lugares. De certa forma, é uma maneira de condensar as informações presentes nos gráficos anteriores em uma única visualização. Boa parte dos locais têm divergência próxima a 0, com poucas exceções significativas para valores positivos ou negativos. A célula com maior divergência foi (-46,4745, -23,5755), no distrito do Parque do Carmo, em São Paulo, com valor igual a 13. A célula com menor divergência foi (-46.3725, -23.5695), no município de Ferraz de Vasconcelos, bairro Vila São Sebastião, cujo valor encontrado foi -13.

Como já mencionado anteriormente, (-46,5405, -23,5935) aparece como a maior interme-

Figura 38 – Distribuição da divergência na rede direcionada



(a) Em todo o estado



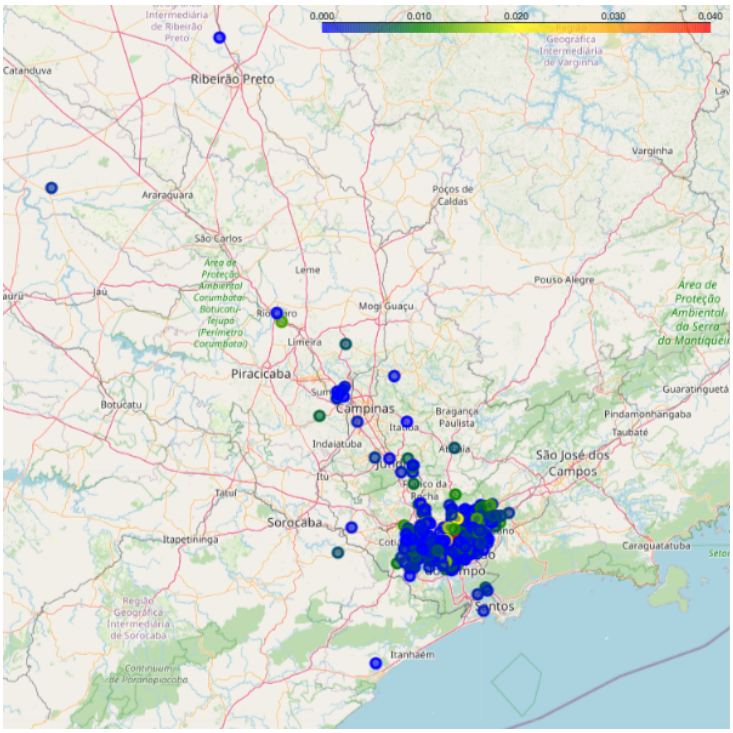
(b) Nas células da capital

Gerada pelo autor

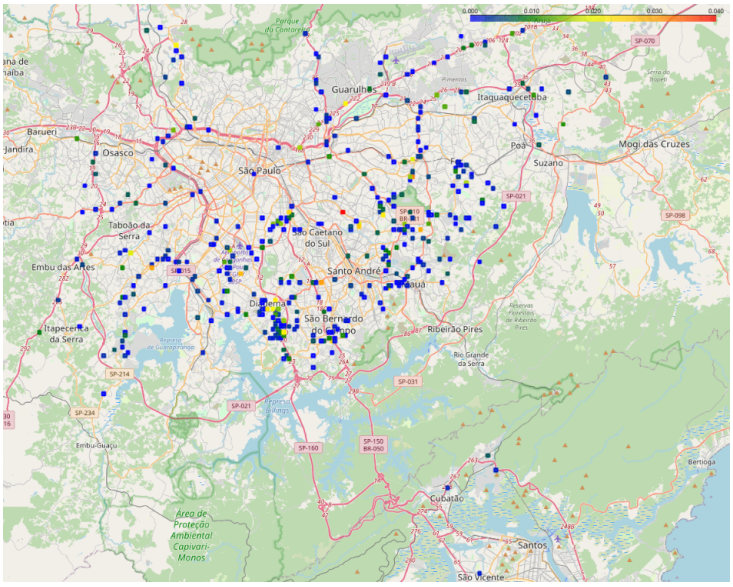
diância na rede direcionada, o que pode ser verificado na Figura 39, sugerindo que eventos nessa célula são frequentemente intermediários para a comunicação entre outras células. No restante das células houve uma distribuição uniforme, com a maior parte dos locais significativamente próxima de zero em relação ao máximo, que foi alcançado novamente pela célula localizada no distrito de São Lucas. Pouquíssimos valores foram altos ou intermediários.

Os coeficientes de agrupamento local de entrada e saída medem a densidade de conexões

Figura 39 – Distribuição da intermediância na rede direcionada



(a) Em todo o estado

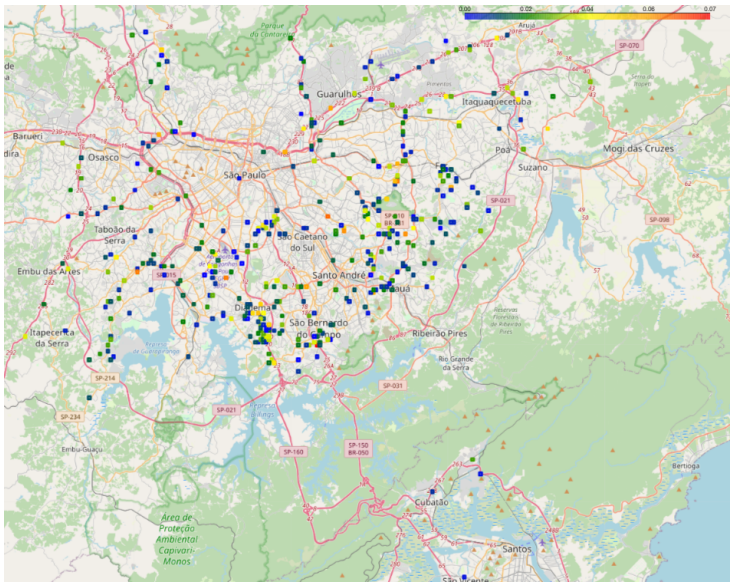


(b) Nas células da capital

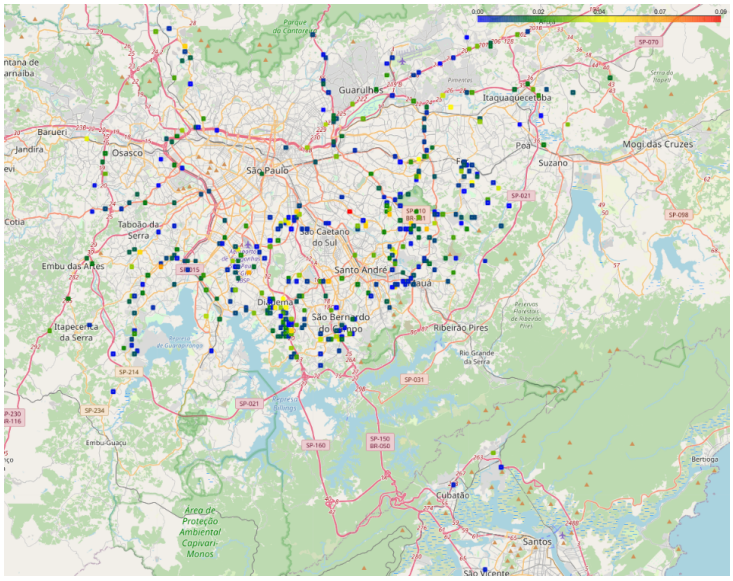
Gerada pelo autor

entre os ”vizinhos de entrada e saída” de uma célula. Como, em um grafo direcionado, formar um subgrafo completo é tarefa bem mais difícil, devido à direção das arestas, o coeficiente de agrupamento tende a ser muito menor também. Na distribuição, presente na Figura 40, foi possível observar que muitas células apresentam coeficientes de agrupamento local de entrada e saída próximos a zero, indicando uma baixa densidade de conexões locais. No entanto, algumas células na capital apresentam valores mais elevados para essas medidas, indicando uma maior

Figura 40 – Distribuição do agrupamento local nas células da capital para rede direcionada



(a) Coeficiente de Agrupamento de entrada



(b) Coeficiente de Agrupamento de saída

Gerada pelo autor

densidade de conexões locais nessas células específicas.

5.2 Conclusão

Ambas as redes revelaram estruturas e valores interessantes, destacando perfis consistentes, especialmente quando submetidas a uma comparação detalhada dentro do contexto e natureza de cada método empregado. A abordagem de redes emerge como uma ferramenta valiosa para a polícia científica, pois estabelece conexões entre diversas regiões do estado com base em seus padrões criminais. Essa abordagem possibilita a identificação de elementos anteriormente ocultos nos dados, como a presença de atividades criminosas organizadas ou áreas particularmente

vulneráveis, uma vez que a sincronização entre os eventos em cada série destaca uma associação entre as ocorrências.

As medidas resultantes da escolha de um perfil específico de *Event Synchronization* para avaliar a similaridade entre as séries de eventos das células demonstraram uma certa afinidade. Isso indica que, de maneira geral, a essência da informação obtida por cada perfil é semelhante. No entanto, uma abordagem direcionada para o *Event Synchronization* parece oferecer vantagens significativas para fins preditivos. Nesse enfoque específico, é possível traçar a cronologia entre a ocorrência de eventos em locais sincronizados, proporcionando a oportunidade de usar os registros de novas ocorrências em um determinado local para antecipar e mitigar delitos em outros locais.

Os resultados apontam para uma influência muito forte da região metropolitana da cidade de São Paulo na dinâmica criminal dos roubos de veículos como um todo. As células que foram filtradas pela quantidade de delitos em si, na Seção 4.4.2, já eram majoritariamente dessa região. Para além disso, os nós mais centrais, em ambas as redes, para quaisquer medidas, estavam situados nela. A constatação de que certas localidades, especialmente aquelas em grandes centros urbanos, exercem influência sobre eventos de roubos de veículos em outras áreas, alinha-se com teorias sociológicas que enfatizam a influência do ambiente densamente urbanizado na ocorrência de crimes (Bulletin d'information sur la criminalité et l'organisation policière, 2000; DURKHEIM; LUKES, 1982; GLAESER; SACERDOTE, 1999). Assim sendo, os resultados terem destacado essa região está dentro das expectativas.

Esses achados respaldam a importância de um enfoque interdisciplinar na compreensão e combate à criminalidade, destacando a necessidade contínua de pesquisas e estudos inovadores na área. A intersecção entre análises estatísticas avançadas, como as realizadas por meio das redes complexas e técnicas específicas de sincronização de eventos, oferece um panorama abrangente e dinâmico dos padrões criminais. O reconhecimento da influência regional na propagação de eventos criminais reforça a pertinência de investigações detalhadas e específicas, ressaltando a importância do ambiente social na ocorrência de crimes. Esse entendimento mais profundo não apenas esclarece a complexidade do fenômeno criminal, mas também abre caminho para estratégias mais precisas e direcionadas por parte das autoridades, permitindo uma abordagem proativa na prevenção dos crimes.

REFERÊNCIAS

ALBERT, R.; BARABáSI, A.-L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, American Physical Society (APS), v. 74, n. 1, p. 47–97, jan. 2002. ISSN 1539-0756.

ALRUHAYMI, A. Z.; KIM, C. J. Study on the missing data mechanisms and imputation methods. **Open Journal of Statistics**, Scientific Research Publishing, Inc., v. 11, n. 04, p. 477–492, 2021. ISSN 2161-7198.

AMISANO, M. De um direito penal antropocêntrico a um direito penal antropomórfico. v. 1, p. 87–103, 2018. ISSN 2595-2935.

BARABáSI, A.-L.; POSFAI, M. **Network Science**. [*S.l.: s.n.*]: Cambridge University Press, 2016.

BARRAT, A. **Dynamical processes on complex networks**. Cambridge: Cambridge University Press, 2013. Title from publisher's bibliographic system (viewed on 05 Oct 2015).

BENESTY, J. *et al.* Pearson correlation coefficient. *In*: **Noise Reduction in Speech Processing**. [*S.l.: s.n.*]: Publisher, 2009. p. 1–4.

BOERS, N. **Complex network analysis of extreme rainfall in South America**. 06 2015. Tese (Doutorado), 06 2015.

BOERS, N. *et al.* Prediction of extreme floods in the eastern central andes based on a complex networks approach. **Nature Communications**, Springer Science and Business Media LLC, v. 5, n. 1, out. 2014. ISSN 2041-1723.

BOTHOS, J. M. A.; THOMOPOULOS, S. C. A. Factors influencing crime rates: an econometric analysis approach. *In*: KADAR, I. (ed.). **SPIE Proceedings**. [*S.l.: s.n.*]: SPIE, 2016. ISSN 0277-786X.

BUCKLEY, A.; BUTLER, K. A. Dealing with missing data. **ARCUSER**, 2017. Summer 2017. Disponível em: https://www.esri.com/about/newsroom/arcuser/dealing-with-missing-data/.

Bulletin d'information sur la criminalité et l'organisation policière. Québec, volume 2, nº 2. decembre 2000.

CHESNAIS, J. C. A violência no brasil: causas e recomendações políticas para a sua prevenção. **Ciência & Saúde Coletiva**, FapUNIFESP (SciELO), v. 4, n. 1, p. 53–69, 1999. ISSN 1413-8123.

COSTA, A. T. M. **Problemas da Investigação de Homicídios no Brasil**. 2021. https://fontesegura.forumseguranca.org.br/problemas-da-investigacao-de-homicidios-no-brasil/. Fontesegura. Brasília. Acesso em: 12 out. 2023.

COUSINEAU, D.; CHARTIER, S. Outliers detection and treatment: A review. **International Journal of Psychological Research**, v. 3, 06 2010.

COVER, T. M.; THOMAS, J. A. **Elements of information theory**. Hoboken, NJ: Wiley-Interscience, 2001. ISBN 9780471200611.

DANIEL, W. **Applied Nonparametric Statistics**. Duxbury, 2000. (Classic Series). ISBN 9780534381943. Disponível em: https://books.google.com.br/books?id=bCDFAAAACAAJ.

DIAS, D. O. Pobreza, criminalidade e direitos sociais: Causas, consequÊncias e possÍveis soluÇÕes. v. 14, p. 53–63, 2019. ISSN 1983-4225.

DIXON, A.; FARRELL, G. Age-period-cohort effects in half a century of motor vehicle theft in the united states. **Crime Science**, v. 9, n. 1, p. 17, 10 2020. ISSN 2193-7680. Disponível em: https://doi.org/10.1186/s40163-020-00126-5.

DONNER, R. V.; WIEDERMANN, M.; DONGES, J. F. Complex network techniques for climatological data analysis. *In*: FRANZKE, C.; O'KANE, T. (ed.). **Complex Systems in the Natural and Social Sciences**. 1. ed. Cambridge: Cambridge University Press, 2017. p. 159–183.

DOROGOVTSEV, S. N. **Lectures on complex networks**. Oxford: Oxford University Press, 2010. (Oxford master series in physics, no. 20). Includes bibliographical references and index.

DURKHEIM, E.; LUKES, S. **Rules of Sociological Method**. Free Press, 1982. (Contemporary social theory). ISBN 9780029079409. Disponível em: https://books.google.com.br/books?id=dM01B9O6s8YC.

FANG, C.; WANG, C. **Time Series Data Imputation: A Survey on Deep Learning Approaches**. [*S.l.: s.n.*]: arXiv, 2020.

FREITAS, J. B.; CLARINDO, J. P.; AGUIAR, C. D. SPSafe: um dataset sobre dados de criminalidade no estado de são paulo. *In*: **Anais do V Dataset Showcase Workshop (DSW 2023)**. [*S.l.: s.n.*]: Sociedade Brasileira de Computação, 2023.

GARLASCHELLI, D.; LOFFREDO, M. I. Patterns of link reciprocity in directed networks. **Physical Review Letters**, American Physical Society (APS), v. 93, n. 26, dez. 2004. ISSN 1079-7114. Disponível em: http://dx.doi.org/10.1103/PhysRevLett.93.268701.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 99, n. 12, p. 7821–7826, jun. 2002. ISSN 1091-6490.

GLAESER, E. L.; SACERDOTE, B. Why is there more crime in cities? **Journal of Political Economy**, The University of Chicago Press, v. 107, n. S6, p. S225–S258, 1999. ISSN 00223808, 1537534X. Disponível em: http://www.jstor.org/stable/10.1086/250109.

GOMES, L. D. *et al.* Crimes na era covid-19: evidências para o estado de são paulo. **Revista Brasileira de Segurança Pública**, Revista Brasileira de Seguranca Publica, v. 17, n. 2, p. 370–393, ago. 2023. ISSN 1981-1659.

GUEDES, C. A ciência no combate ao crime. **Editora JC**, setembro 30 2007. 2º Vice-Presidente do TJ/RJ. Disponível em: https://www.editorajc.com.br/a-ciencia-no-combate-ao-crime/. Acesso em: 1 dez 2023.

HAGE, P.; HARARY, F. Eccentricity and centrality in networks. **Social Networks**, v. 17, n. 1, p. 57–63, 1995. ISSN 0378-8733. Disponível em: https://www.sciencedirect.com/science/article/pii/0378873394002489.

HAINES, K. Crime is a social problem. **European Journal on Criminal Policy and Research**, Springer Science and Business Media LLC, v. 7, n. 2, p. 263–275, 1999. ISSN 0928-1371.

HYUN, K. The prevention and handling of the missing data. **Korean J Anesthesiol**, v. 64, n. 5, p. 402–406, 2013. Disponível em: http://ekja.org/journal/view.php?number=7569.

Instituto de Pesquisa Econômica Aplicada (Ipea). **Atlas da Violência - Taxa de Homicídios**. 2022. Consultado em 4 de Dezembro de 2023. Disponível em: https://www.ipea.gov.br/atlasviolencia/publicacoes.

JUNIOR, M. de O. M. **O PAPEL DO DIREITO PENAL, PROCESSO PENAL E DA LEI DE EXECUÇÃO PENAL NO SISTEMA PUNITIVO**. 2021. 15-16 p.

KENDALL, M.; GIBBONS, J. **Rank Correlation Methods**. Edward Arnold, 1990. (A Charles Griffin title). ISBN 9780852643051. Disponível em: https://books.google.com.br/books?id=ly4nAQAAIAAJ.

KENDALL, M. G. A new measure of rank correlation. **Biometrika**, v. 30, n. 1-2, p. 81–93, 06 1938. ISSN 0006-3444. Disponível em: https://doi.org/10.1093/biomet/30.1-2.81.

KEOGH, E. *et al.* Dimensionality reduction for fast similarity search in large time series databases. **Knowledge and Information Systems**, v. 3, p. 263–286, 2001.

KUMAR, S. **ML Algorithm that Natively Supports Missing Values**. 2022. Published in Towards Data Science, Jan 18, 2022. Disponível em: https://towardsdatascience.com/ml-algorithm-that-natively-supports-missing-values-40b42559c1ec.

MAGALHÃES, M. **Probabilidade e Variáveis Aleatórias**. Edusp, 2006. ISBN 9788531409455. Disponível em: https://books.google.com.br/books?id=Pel8ATx9QDQC.

MARCHEZINI, B. R.; SPOLADOR, H. F. S.; JORGE, M. A. Crescimento econômico e criminalidade: uma análise de dados em painel para o estado de são paulo. **Associação Brasileira de Estudos Regionais**, v. 6, n. 2, p. 178–197, 2020. Disponível em: https://brsa.org.br/wp-content/uploads/wpcf7-submissions/6840/Crescimento-e-Criminalidade-artigo-id.pdf.

MIOT, H. A. Valores anômalos e dados faltantes em estudos clínicos e experimentais. **Jornal Vascular Brasileiro**, FapUNIFESP (SciELO), v. 18, 2019. ISSN 1677-5449.

MORETTIN, C. M. C. T. P. A. **Análise de Séries Temporais**: Modelos lineares univariados. [*S.l.: s.n.*]: Blucher, 2006. ISBN 9788521213512.

NAGAMOCHI, H.; IBARAKI, T. **Algorithmic Aspects of Graph Connectivity**. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, Sao Paulo, Delhi: Cambridge University Press, 2008. First published 2008. ISBN 978-0-521-87864-7.

NEWMAN, M. E. J. **Networks: An Introduction**. Oxford; New York: Oxford University Press, 2010. ISBN 9780199206650 0199206651. Disponível em: http://www.amazon.com/Networks-An-Introduction-Mark-Newman/dp/0199206651/ref=sr_1_5?ie=UTF8&qid=1352896678&sr=8-5&keywords=complex+networks.

ODENWELLER, A.; DONNER, R. V. Disentangling synchrony from serial dependency in paired-event time series. **Physical Review E**, American Physical Society (APS), v. 101, n. 5, p. 052213, may 2020.

PAUW, E. D. (ed.). **Technology-led policing**. Antwerpen [u.a.]: Maklu, 2011. (Cahiers politiestudies, 20). ISBN 9789046604120.

Portal do Governo do Estado de São Paulo. **Polícia Civil mira na receptação e desmanche de veículos**. 2012. Acesso em: 01 dez. 2023. Disponível em: https://www.saopaulo.sp.gov.br/ultimas-noticias/policia-civil-mira-na-receptacao-e-desmanche-de-veiculos-1/.

QUIROGA, R. Q.; KREUZ, T.; GRASSBERGER, P. Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. **Physical Review E**, American Physical Society (APS), v. 66, n. 4, p. 041904, oct 2002.

RHEINWALT, A. *et al.* Boundary effects in network measures of spatially embedded networks. **EPL (Europhysics Letters)**, IOP Publishing, v. 100, n. 2, p. 28002, out. 2012. ISSN 1286-4854.

RUITER, S. Crime location choice. **The Oxford handbook of offender decision making**, Oxford University Press Oxford, p. 398–420, 2017.

SAMPAIO, H. V. P. A. **Introdução à criminologia**. [*S.l.: s.n.*], 2023.

SANTOS, J. C. dos. **A Criminologia da Repressão: Crítica à Criminologia Positivista**. [*S.l.: s.n.*]: Tirant Brasil, 2019.

SERRà, J.; ARCOS, J. L. An empirical evaluation of similarity measures for time series classification. arXiv, 2014.

SHADBAHR, T. *et al.* The impact of imputation quality on machine learning classifiers for datasets with missing values. **Communications Medicine**, Springer Science and Business Media LLC, v. 3, n. 1, out. 2023. ISSN 2730-664X.

SHUMWAY, R. H.; STOFFER, D. S. (ed.). **Time Series Analysis and Its Applications**: With r examples. New York, NY: Springer Science+Business Media, LLC, 2006. (SpringerLink). Includes bibliographical references and index.

SLEEUWEN, S. E. M. van; STEENBEEK, W.; RUITER, S. When do offenders commit crime? an analysis of temporal consistency in individual offending patterns. **Journal of Quantitative Criminology**, Springer Science and Business Media LLC, v. 37, n. 4, p. 863–889, ago. 2020. ISSN 1573-7799.

TAMBOLI, N. **Effective Strategies for Handling Missing Values in Data Analysis**. 2023. Updated on July 14th, 2023. Disponível em: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/#:~:text=If%20you%20are%20aiming%20for,the%20accuracy%20of%20the%20model.

WALLACE, M. **Exploring the involvement of organized crime in motor vehicle theft**. Ottawa: Statistics Canada, Canadian Centre for Justice Statistics, 2004. Note de reconnaissance: Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. ISBN 0-660-19309-4. Disponível em: https://www150.statcan.gc.ca/n1/en/pub/85-563-x/85-563-x2004001-eng.pdf?st=Qev8Tuxw.

WALSH, J. A.; TAYLOR, R. B. Predicting decade-long changes in community motor vehicle theft rates: Impacts of structure and surround. **Journal of Research in Crime and Delinquency**, SAGE Publications, v. 44, n. 1, p. 64–90, fev. 2007. ISSN 1552-731X.

WAWRZYNIAK, Z. M. *et al.* Relationships between crime and everyday factors. *In*: **2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)**. [*S.l.: s.n.*]: IEEE, 2018.

WEI, W. W. S. **Time series analysis**: Univariate and multivariate methods. Reprint. with corr. Redwood City, Calif. [u.a.]: Addison-Wesley, 1994. (The @advanced book programm). ISBN 0201159112.

WOLF, F. P. W. **Complex networks across fields: from climate variability to online dynamics**. [*S.l.: s.n.*]: Humboldt-Universität zu Berlin, 2021.

APÊNDICES

APÊNDICE A – TABELA DOS DADOS DE VEÍCULOS

O Quadro 2 possui todas as colunas presentes nas planilhas de roubo e furto de veículos disponíveis no site da transparência da SSP-SP.

Quadro 2 – Campos contidos nas tabelas de roubo de veículo

Campo	**Descrição**	**Tipo de Dado (ideal)**
ANO_BO	Ano do BO	inteiro
NUM_BO	Número identificador do BO	Inteiro
BO_INICIADO	Data de início do BO	Data
BO_EMITIDO	Data de conclusão do BO	Data
DATAOCORRENCIA	Data da ocorrência	Data
HORAOCORRENCIA	Hora da ocorrência	Hora
PERIODOOCORRENCIA	Período do dia da ocorrência	Texto
DATACOMUNICACAO	Data de comunicação	Data
DATAELABORACAO	Data de elaboração	Data
BO_AUTORIA	Autoria do BO	Texto
FLAGRANTE	Se houve flagrante	Booleano
LOGRADOURO	Logradouro da ocorrência	Texto
NUMERO	Número da ocorrência	Texto
BAIRRO	Bairro da ocorrência	Texto
CIDADE	Cidade da ocorrência	Texto
UF	Unidade federativa	Texto
LATITUDE	Latitude da ocorrência	Ponto flutuante
LONGITUDE	Longitude da ocorrência	Ponto flutuante
DESCRICAOLOCAL	Descrição do local	Texto
EXAME	Exame de perícia	Texto
SOLUCAO	Solução	Texto
DELEGACIA_NOME	Nome da delegacia	Texto
DELEGACIA_CIRCUNSCRICAO	Delegacia de circunscrição	Texto
ESPECIE	Espécie	Texto
RUBRICA	Rubrica	Texto
DESDOBRAMENTO	Desdobramento	Texto
STATUS	Status da ocorrência	Texto
TIPOPESSOA	Responsável pelo BO	Texto
VITIMAFATAL	Se houve vítima Fatal	Booleano
NACIONALIDADE	Nacionalidade da vítima	Texto
SEXO	Sexo da vítima	Texto
DATANASCIMENTO	Data de nascimento da vítima	Data
IDADE	Idade da vítima	Inteiro
PROFISSAO	Profissão da vítima	Texto
GRAUINSTRUCAO	Grau de Instrução da vítima	Texto
CORCUTIS	Cor da Pele da vítima	Texto
NATUREZAVINCULADA	Natureza Vinculada do delito	Texto
PLACA_VEICULO	Placa do veículo	Texto
UF_VEICULO	Unidade federativa do veículo	Texto
CIDADE_VEICULO	Cidade do veículo	Texto
DESCR_COR_VEICULO	Cor do veículo	Texto
DESCR_MARCA_VEICULO	Marca do veículo	Texto
ANO_FABRICACAO	Ano de Fabricação	Inteiro
ANO_MODELO	Ano do Modelo	Inteiro
DESCR_TIPO_VEICULO	Tipo do Veículo	Texto

Fonte: Elaborado pelo autor.

APÊNDICE B – TRATAMENTOS DOS DADOS

Em qualquer base de dados extensa e representativa do mundo real, é inevitável encontrar valores ausentes, inconsistências e outros problemas de integridade (SHADBAHR *et al.*, 2023). Os valores ausentes podem surgir devido a diversos motivos, como falhas na coleta, erros de digitação ou mesmo pela própria natureza do fenômeno estudado. Da mesma forma, inconsistências podem ocorrer devido a processos de coleta de dados não padronizados, falta de controle de qualidade ou até mesmo por conta de mudanças nos procedimentos ao longo do tempo (MIOT, 2019). Nesse contexto, ao lidar com bases de dados extensas, é desejável empregar técnicas de limpeza e pré-processamento para assegurar a qualidade e confiabilidade das informações que serão utilizadas em análises e tomadas de decisão (ALRUHAYMI; KIM, 2021).

B.1 Por que tratar valores presentes?

Valores presentes na base de dados podem, por vezes, apresentar uma série de problemas, seja devido à falta de padronização na aferição, equívocos durante o preenchimento ou até mesmo devido a variações inesperadas nos métodos de coleta, entre outras inconsistências. Essas divergências podem resultar em um desafio significativo no que diz respeito à análise e interpretação dos dados, uma vez que, quando se trata de grandes volumes de informações, realizar uma análise minuciosa de cada entrada torna-se impraticável e inviável.

O primeiro passo para tratar valores inconsistentes é definir qual tipo de entrada esperada para cada campo na base de dados. Ao definir o tipo de entrada esperada, são estabelecidos critérios claros sobre o formato e o intervalo ou conjunto aceitável de valores. Com essa base, é possível identificar e corrigir inconsistências de forma mais eficaz.

Em seguida, é preciso estabelecer uma forma para substituição dos valores presentes. Se eles podem ser substituídos por um equivalente (que padronize entradas diferentes que representem a mesma informação) ou se o valor em questão trata-se de um *outlier*. Neste último caso, é importante avaliar se a remoção desse valor é a abordagem mais apropriada. Em alguns casos, *outliers* podem fornecer informações valiosas sobre o conjunto de dados, mas em outros, sua presença pode distorcer análises estatísticas (COUSINEAU; CHARTIER, 2010).

Além disso, é fundamental documentar todas as decisões tomadas durante o processo de tratamento de dados, incluindo a definição de critérios, as escolhas de substituição ou remoção de valores, e qualquer transformação aplicada. Por fim, é recomendável realizar verificações adicionais após o tratamento dos dados para garantir que as inconsistências foram adequadamente abordadas e que o conjunto de dados está pronto para análises mais aprofundadas.

B.1.1 Caminhos para tratar valores presentes nos dados da SSP-SP

Um dos campos cruciais para este estudo, que demandava especial atenção quanto aos valores inconsistentes, foi o 'BAIRRO'. Durante a elaboração dos boletins, este campo frequentemente carece de preenchimento adequado, sendo por vezes associado a pontos de

referência ou nomes de ruas ou parques. Além disso, a nomenclatura de alguns bairros pode variar, apresentando mais de uma denominação popular. Uma abordagem para tratar esses valores é empregar as informações de 'LATITUDE' e 'LONGITUDE' para identificar o bairro da ocorrência de maneira direta. Recomenda-se a utilização de um *shapefile* contendo as delimitações dos bairros na região de interesse, o qual pode ser acessado através do site oficial da prefeitura da respectiva cidade.

Para o presente trabalho, apenas as instâncias das cidades de São Paulo, São Bernardo e Guarulhos tiveram o tal campo tratado. Inicialmente, pretendia-se tratar apenas as 5 cidades com mais instâncias, já que o processo para todas as cidades de registro seria exaustivo e pouco influenciaria o resultado final, mas Campinas e Santo André não possuíam *shapefiles*, ou qualquer outro tipo de documento que permitisse determinar precisamente os bairros a partir das coordenadas geográficas, nos sites das suas prefeituras.

B.2 Por que tratar valores ausentes?

Valores ausentes na base de dados representam um problema relevante, não só para a análise, quando o percentual de valores ausentes não é desprezível em relação ao todo, como também para o emprego de métodos estatísticos ou computacionais que geram tomada de decisão. A presença de lacunas na informação pode comprometer a integridade dos resultados obtidos, levando a conclusões precipitadas ou imprecisas (TAMBOLI, 2023). Além disso, para algoritmos e técnicas que dependem fortemente dos dados disponíveis, a ausência de informações pode inviabilizar o processo, resultando em escolhas subótimas ou até mesmo incorretas. De fato, a maior parte dos algoritmos modernos de aprendizado de máquina não consegue gerar resultados em cima de dados com valores ausentes (KUMAR, 2022).

O primeiro passo para tratar valores ausentes é identificá-los. Existem diversas maneiras de identificar valores ausentes em um conjunto de dados. Uma abordagem simples é utilizar métodos estatísticos descritivos, como a contagem de valores nulos em cada coluna ou a porcentagem de valores ausentes em relação ao total de observações.

Após identificar os valores ausentes, a determinação da estratégia de tratamento torna-se intrinsicamente ligada ao contexto do problema e à natureza dos dados. A análise do contexto envolve considerações específicas, como a presença de dados geoespaciais ou temporais, que podem impactar diretamente na escolha da abordagem adequada. Por exemplo, em dados geoespaciais, a proximidade física entre observações pode ser crucial (BUCKLEY; BUTLER, 2017), enquanto em dados temporais, padrões ao longo do tempo podem influenciar as decisões de imputação (FANG; WANG, 2020).

Além disso, a natureza dos dados desempenha um papel fundamental na seleção da estratégia mais apropriada. Se as informações presentes permitem inferir de maneira confiável os valores ausentes, a imputação pode ser uma opção viável. Abordagens comuns incluem a exclusão de observações ou colunas com valores ausentes, a imputação utilizando estatísticas descritivas, como média, mediana ou moda, e métodos mais avançados, como a imputação baseada em modelos estatísticos ou algoritmos de aprendizado de máquina.

É importante ressaltar que a decisão sobre como lidar com os valores ausentes deve ser fundamentada em uma compreensão profunda do conjunto de dados e no impacto potencial nas análises subsequentes. Cada abordagem possui suas vantagens e limitações, e a escolha dependerá das características específicas do problema em questão. A transparência e documentação adequada das etapas de tratamento de valores ausentes são essenciais para garantir a reprodutibilidade e a validade das análises realizadas.

B.2.1 Caminhos para tratar valores ausentes nos dados da SSP-SP

Como discutido na Seção 4.2.1, a abordagem adotada neste trabalho para lidar com valores ausentes nos campos listados no Quadro 1 foi a exclusão das respectivas instâncias. Após a remoção em torno dos campos 'LATITUDE', 'LONGITUDE' e 'HORAOCORRENCIA', não restaram quaisquer valores ausentes para os demais campos.

Devido ao fato de o número de instâncias com valores ausentes não ser significativo em relação ao conjunto de dados como um todo, e considerando que não há razão para acreditar que essas omissões ocorram de maneira aleatória[1], a opção pela exclusão foi considerada uma escolha razoável, embora essa abordagem introduza um leve viés na análise final.

Uma segunda abordagem considerada, embora não tenha sido implementada, é a imputação. Existem diversas metodologias e algoritmos disponíveis para realizar esse processo. No entanto, devido à natureza da imputação, que opera sobre valores ausentes, determinar com precisão o quão bem-sucedida ela foi torna-se desafiador, sendo possível apenas realizar estimativas com base em hipóteses. Para mais informações sobre a técnica, recomenda-se Shadbahr *et al.* (2023), Hyun (2013)

Para os campos de latitude e longitude, foram conduzidos testes utilizando regressores 'linear', 'KNN' (com $N = 3$), e 'Árvore de Decisão', além de uma abordagem de imputação baseada em informações de bairro e cidade. As variáveis independentes em todas as regressões foram 'CIDADE' e 'BAIRRO', que passaram por um processo de *encoding*. A Tabela 10 traz os resultados, em grau, dos modelos para as principais métricas no contexto de regressão.

Tabela 10 – Resultados de avaliação dos modelos para 'LATITUDE' E 'LONGITUDE'

Modelo	**Erro médio Absoluto**	**Erro Quadrático Médio**	**Coeficiente de Determinação**
KNN	0.085	0.101	0.653
Linear	0.085	0.101	0.653
Árvore de Decisão	0.013	0.013	0.955
Bairro e Cidade	0.007	0.000	0.999

Fonte: Elaborada pelo autor.

As métricas foram calculadas mediante a divisão dos dados, excluindo aqueles com valores ausentes, em conjuntos de treinamento e teste. Portanto, representam uma avaliação do quão eficazes foram as imputações dentro do contexto dos dados conhecidos. Para generalizar

[1] há várias regiões em que a polícia teria dificuldade de adentrar para aferir a latitude e a longitude, por exemplo.

conclusões sobre a utilização dessas imputações nos valores verdadeiramente ausentes, é necessário pressupor que os dados ausentes se comportam de maneira semelhante aos dados presentes.

Destaca-se que a estratégia de imputação baseada em bairro e cidade demonstrou ser a mais bem-sucedida. Essa abordagem envolve a imputação dos valores ausentes utilizando a mediana das instâncias com valores presentes que compartilham o mesmo bairro e cidade. Caso não haja correspondência para o par cidade e bairro, a imputação é realizada com base apenas na cidade. No caso de ausência de correspondentes, nenhuma imputação é realizada, indicando que, ao contrário das regressões convencionais, essa estratégia não fornece respostas para todos os valores ausentes, o que é uma desvantagem relevante em muitos casos.

O mesmo procedimento pode ser usado para imputar valores em 'HORAOCORRENCIA', a partir de 'CIDADE', 'BAIRRO' e 'PERIODOOCORRENCIA'. Porém, não é possível usar a categoria 'EM HORA INCERTA', presente em 'PERIODOOCORRENCIA'. Por motivos de ajuste, os testes foram feitos utilizando-se regressores 'KNN' (com N=3), 'Árvore de Decisão' e 'Random Forest', conforme Tabela 11, cujas métricas estão expostas em minutos.

Tabela 11 – Resultados de avaliação dos modelos para 'HORAOCORRENCIA'

Modelo	**Erro Médio Absoluto**	**Erro Quadrático Médio**	**Coeficiente de Determinação**
KNN	150.074	52723.364	0.687
Árvore de Decisão	84.617	11292.0194	0.933
Random Forest	83.141	10242.456	0.939

Fonte: Elaborada pelo autor.

O regressor 'Random Forest' se mostrou mais eficiente do que os demais, levemente superior a 'Árvore de Decisão'. Contudo, o erro cometido por qualquer um dos algoritmos apresentados não parece ser desprezível numa análise que se pretenda criteriosa. Todos os regressores empregados basearam-se nas implementações correspondentes disponíveis na biblioteca scikit-learn[2], que oferece uma ampla variedade de algoritmos e ferramentas computacionais adicionais capazes de auxiliar na execução da tarefa.

—

[2] https://scikit-learn.org/