

# UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

**Efeito de congestionamentos em acidentes de trânsito: estudo de caso de São Paulo**

**Pedro Gigeck Freire**

Monografia - MBA em Inteligência Artificial e Big Data



SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: \_\_\_\_\_

**Pedro Gigeck Freire**

## **Efeito de congestionamentos em acidentes de trânsito: estudo de caso de São Paulo**

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Prof. Dr. Caetano Mazzoni Ranieri

**Versão original**

**São Carlos  
2024**

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO,  
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E  
PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

FREIRE, Pedro Gigeck

Efeito de congestionamentos em acidentes de trânsito: estudo de caso de São Paulo / Pedro Gigeck Freire ; orientador: Prof. Dr. Caetano Mazzoni Ranieri. – São Carlos, 2024.

54 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Ranieri, Caetano Mazzoni, orient. II. Título.

**Pedro Gigeck Freire**

## **Effect of congestion on traffic accidents: São Paulo case study**

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Caetano Mazzoni Ranieri

**Original version**

**São Carlos  
2024**



*“Enxergar os eventos por uma bola de cristal é possível, mas, infelizmente, apenas depois que eles já aconteceram. (...) Depois da onça morta, todo mundo é caçador.”*

*Leonard Mlodinow*



## **RESUMO**

**FREIRE, P. G. Efeito de congestionamentos em acidentes de trânsito: estudo de caso de São Paulo.** 2024. 54 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Acidentes de trânsito e congestionamentos são fenômenos urbanos que geram significativos impactos sociais e econômicos. Este trabalho investiga o uso de algoritmos de agrupamento e regressão para identificar a correlação entre esses temas. Foram analisados dados da cidade de São Paulo por meio de um processamento georreferenciado em estrutura de grafo, permitindo visualizações em formato de mapas e diagramas. Comparamos diferentes algoritmos para agrupar os registros de acidentes e congestionamentos, destacando o DBSCAN como a melhor estratégia para dados georreferenciados. Além disso, os modelos de regressão revelaram uma correlação inversa entre a gravidade dos acidentes e a intensidade dos congestionamentos, sugerindo que a lentidão dos veículos pode mitigar a severidade dos incidentes. O coeficiente de determinação  $R^2$  dos modelos é de cerca de 40%, indicando oportunidades para a inclusão de outras fontes de dados no estudo. Este trabalho sugere ações baseadas nos dados investigados e em revisão de literatura a redução da gravidade dos acidentes veiculares. Tecnologias como a ciência de dados e a inteligência artificial demonstram um grande potencial para transformar a segurança viária e reduzir o número de vidas perdidas.

**Palavras-chave:** Acidente de Trânsito. Congestionamento. Inteligência Artificial. Agrupamento. Regressão.



## **ABSTRACT**

FREIRE, P. G. **Effect of congestion on traffic accidents: São Paulo case study.** 2024. 54 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Traffic accidents and congestion are urban phenomena that generate significant social and economic impacts. This work investigates the use of clustering and regression algorithms to identify the correlation between these issues. We analyzed data from the city of São Paulo through georeferenced processing in a graph structure, allowing visualizations in the form of maps and diagrams. We compared different algorithms to cluster the records of accidents and congestion, highlighting DBSCAN as the best strategy for georeferenced data. Furthermore, regression models revealed an inverse correlation between the severity of accidents and the intensity of congestion, suggesting that vehicle slowdowns may mitigate the severity of accidents. The coefficient of determination  $R^2$  of the models is around 40%, indicating opportunities for including other data sources in the study. This work suggests actions based on the investigated data and literature review to reduce the severity of vehicular accidents. Technologies such as data science and artificial intelligence demonstrate great potential to transform road safety and reduce the number of lives lost.

**Keywords:** Traffic Accidents. Congestion. Artificial Intelligence. Clustering. Regression



## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>15</b>
1.1	<b>Justificativa e importância</b>	<b>16</b>
1.2	<b>Objetivos</b>	<b>16</b>
1.3	<b>Considerações Finais</b>	<b>17</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>19</b>
2.1	<b>Mobilidade Urbana</b>	<b>19</b>
2.1.1	Acidentes de Trânsito	19
2.1.2	Congestionamentos	20
2.2	<b>Inteligência Artificial</b>	<b>20</b>
2.2.1	Agrupamento de Dados	21
2.2.2	Régressão	23
2.2.3	Métricas de Avaliação	25
2.2.3.1	Métricas de Avaliação para Agrupamentos	25
2.2.3.2	Métricas de Avaliação para Régressão	26
2.3	<b>Trabalhos Relacionados</b>	<b>28</b>
<b>3</b>	<b>METODOLOGIA</b>	<b>31</b>
3.1	<b>Pré-Processamento</b>	<b>31</b>
3.1.1	Vias	32
3.1.2	Acidentes	33
3.1.3	Congestionamentos	35
3.2	<b>Identificação de Arestas</b>	<b>35</b>
3.3	<b>Agrupamento dos dados</b>	<b>38</b>
3.3.1	Eventos	39
3.4	<b>Régressão</b>	<b>40</b>
<b>4</b>	<b>ANÁLISE DOS RESULTADOS</b>	<b>41</b>
4.1	<b>Algoritmos dos Agrupamentos</b>	<b>41</b>
4.2	<b>Características dos Eventos</b>	<b>43</b>
4.3	<b>Modelos de Régressão</b>	<b>45</b>
<b>5</b>	<b>CONCLUSÕES</b>	<b>49</b>
	<b>REFERÊNCIAS</b>	<b>51</b>



## 1 INTRODUÇÃO

Nos últimos séculos, o crescimento dos grandes centros urbanos revolucionou a forma como as pessoas se deslocam no espaço. Durante o século XX, o desenvolvimento das cidades frequentemente priorizou a mobilidade pautada em veículos motorizados (carros, motocicletas, ônibus), promovendo a construção de complexas malhas rodoviárias para sustentar a crescente demanda de carros (Sheller; Urry, 2000; Malatesta, 2014). Nesse contexto, alguns fenômenos negativos relacionados ao tráfego automobilístico, como os congestionamentos, colisões e atropelamentos, surgiram e são cada vez mais comuns em algumas localidades.

Acidentes de trânsito - com ou sem vítimas fatais - são um problema para as cidades, com impactos econômicos e sociais. Em São Paulo, ocorrem cerca de 1000 mortes por ano, com uma tendência de alta (Lara, 2024). Na Austrália, estima-se que os custos de acidentes de trânsito chegam a 27 bilhões de dólares por ano, considerando impactos individuais para as vítimas e também para o estado e empresas (BITRE, 2014). Isso posto, há interesse de autoridades, pesquisadores e da sociedade civil e privada em estudar quais são as causas dessas ocorrências e como evitá-las.

Existem diversas variáveis que podem influenciar na ocorrência de acidentes, como fatores humanos (ingestão de álcool, distração, imprudência, hábitos culturais), e causas extrínsecas aos motoristas, como o clima, qualidade das vias, falhas mecânicas, iluminação e a intensidade do tráfego de veículos (Chand; Jayesh; Bhasi, 2021). Para algumas dessas variáveis, existe uma sólida base na literatura determinando sua relação de causalidade com acidentes, como o consumo de álcool (Martin *et al.*, 2017). Em contrapartida, existem alguns fatores cuja influência na ocorrência de sinistros de trânsito não é consenso entre os pesquisadores. É o caso dos congestionamentos. Em décadas de pesquisa, diversas iniciativas propuseram modelos diferentes para o tema - desde funções lineares positivas até curvas em formas de sino ou 'U' - dependendo de quais variáveis são consideradas nos estudos e qual a granularidade dos dados disponíveis (González; Bedoya-Maya; Calatayud, 2021; Retallack; Ostendorf, 2019).

Este trabalho visa utilizar técnicas de Inteligência Artificial (IA) em conjuntos de dados (datasets) sobre o trânsito da cidade de São Paulo para mapear a relação entre os congestionamentos e os acidentes. Dessa forma, serão comparados os resultados obtidos com a literatura para corroborar teses existentes e apresentados novos dados para tais questões ainda sem consenso científico. Para atingir os objetivos propostos, reunimos bases de dados disponíveis publicamente sobre o tema - particularmente a base sobre lentidão no trânsito de São Paulo da Companhia de Engenharia de Tráfego (CET)<sup>1</sup> e os dados da

---

<sup>1</sup> <http://dados.prefeitura.sp.gov.br/dataset/base-de-dados-sobre-lentidao-por-trechos-cet>

ferramenta Infosiga sobre acidentes de trânsito no estado<sup>2</sup>. Nesses datasets, aplicaremos um processamento para que todas as fontes estejam compatíveis e interpretáveis na semântica geográfica que representam. Ainda em termos de pré-processamento, exercitaremos técnicas de visualização dos dados para entender as distribuições dos registros e ajudar na interpretabilidade dos resultados.

Após o tratamento dos dados, serão aplicados modelos de aprendizado de máquina (a serem detalhados posteriormente) para identificar padrões nos dados e correlações entre as variáveis de interesse e as independentes. Concomitantemente, a revisão da literatura acompanhará o desenvolvimento do trabalho, a fim de validar suposições, comparar estudos com propósitos semelhantes e sugerir técnicas e modelos.

## 1.1 Justificativa e importância

Com as análises desenvolvidas neste trabalho, será possível compreender melhor as relações entre a intensidade do trânsito e seus eventos indesejados, além de contribuir para a identificação de fatores secundários na ocorrência de acidentes. Adicionalmente, comparando resultados de São Paulo com outros estudos feitos em outras regiões, podemos identificar como fatores regionais influenciam nos acidentes, como a cultura local, topografia e distribuição das vias. Da mesma forma, será possível extrair conhecimento de fontes de dados volumosas e complexas sobre acidentes e congestionamentos, através de técnicas de IA, e comparar as conclusões com outros estudos e relatórios técnicos que analisam as variáveis separadamente para a mesma região estudada.

Através das investigações propostas, autoridades de trânsito podem desenvolver políticas baseadas em dados (*data driven*) para tratar o problema com mais assertividade, somando técnicas modernas de IA às experiências de especialistas.

## 1.2 Objetivos

O objetivo geral da pesquisa é determinar a correlação entre congestionamentos de veículos e acidentes de trânsito na cidade de São Paulo. Com base nesse objetivo, visamos realizar os seguintes objetivos específicos:

- Unir dados de congestionamentos e acidentes em uma estrutura adequada que represente as dimensões espacial e temporal do problema;
- Aplicar técnicas de IA para agrupar (clusterizar) os fenômenos de trânsito, identificando padrões nos registros;
- Criar visualizações (mapas) para ilustrar os padrões identificados;

---

<sup>2</sup> <http://catalogo.governoaberto.sp.gov.br/dataset/infosiga-sp-sistema-de-informacoes-gerenciais-de-acidentes-de-transito-do-estado-de-sao-paulo>

- Comparar se as teses propostas por trabalhos relacionados correspondem ao caso de São Paulo;

### **1.3 Considerações Finais**

Ao final deste trabalho, almejamos ter contribuído para a questão - ainda em aberto - da identificação da relevância dos congestionamentos na ocorrência de sinistros de trânsito. É esperado que os algoritmos de Inteligência Artificial utilizados consigam identificar padrões e estatísticas que esclareçam a relação entre os dois temas - e que consigamos traduzir tais resultados para visualizações simples e interpretáveis. Dessa forma, disponibilizaremos insumos para ações bem fundamentadas para combate aos acidentes de trânsito, reforçando a importância da coleta e disponibilidade de dados e oportunizando o estabelecimento de cidades cada vez mais inteligentes<sup>3</sup>.

---

<sup>3</sup> O termo Cidade Inteligente (Smart City) tem sido usado para referenciar abordagens de uso sustentável dos recursos de uma cidade com base em tecnologias e teorias modernas (Batista *et al.*, 2016)



## 2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, abordaremos os conceitos e técnicas que fundamentaram o desenvolvimento desta pesquisa. As bases teóricas aqui discutidas permeiam dois campos de conhecimento principais - Mobilidade Urbana e Inteligencia Artificial - e permitirão conectar com maior precisão as metodologias de processamento com os conteúdos dos conjuntos de dados.

### 2.1 Mobilidade Urbana

Segundo a Política Nacional de Mobilidade Urbana (PNMU), a **mobilidade urbana** é a “condição em que se realizam os deslocamentos de pessoas e cargas no espaço urbano” (Brasil, 2012). Nesse contexto, detalharemos aqui algumas terminologias referentes à mobilidade que serão usadas no decorrer deste trabalho.

#### 2.1.1 Acidentes de Trânsito

Adotando a definição do Código Brasileiro de Trânsito (CTB), “considera-se **trânsito** a utilização das vias por pessoas, veículos e animais, (...) para fins de circulação, parada, estacionamento e operação de carga ou descarga” (Brasil, 1997). Concomitantemente, o termo **Acidente de Trânsito** é popularmente consagrado para nomear diversos tipos de eventos que prejudicam o trânsito de alguma forma. Embora a expressão “acidente de trânsito” também seja usada em contextos técnicos e formais, há uma conotação de involuntariedade no evento. Dessa forma, em 2020, a Associação Brasileira de Normas Técnicas (ABNT) estabeleceu o termo **Sinistro de Trânsito** como

*“todo evento que resulte em dano ao veículo ou à sua carga e/ou em lesões a pessoas e/ou animais, e que possa trazer dano material ou prejuízos ao trânsito, à via ou ao meio ambiente, em que pelo menos uma das partes está em movimento nas vias terrestres ou em áreas abertas ao público.”* (ABNT, 2020)

Neste trabalho, utilizamos “acidentes” e “sinistros” de trânsito como sinônimos.

Ainda segundo a ABNT, podemos ordenar os tipos de eventos pela gravidade do fato, de acordo com a classificação abaixo:

1. Incidentes de trânsito (quando não há vítimas nem dano material, apenas prejuízos ao trânsito);
2. Sinistros de trânsito sem vítimas;

3. Sinistros de trânsito com vítima não-fatal;
4. Sinistros de trânsito com vítima fatal (nesse caso, o óbito pode ocorrer imediatamente ou em até 30 dias da data do acidente).

Utilizamos essa classificação para montar o modelo que correlaciona os sinistros aos congestionamentos na seção 3.

#### 2.1.2 Congestionamentos

O uso da expressão “**congestionamento**” é bastante amplo, dentro e fora do contexto de mobilidade urbana. Aqui, limitaremos o termo com o significado atribuído por Goodwin (2004):

*“Congestionamento é definido como a dificuldade que veículos impõem uns aos outros, devido a relação entre o fluxo (de veículos) e velocidade, em condições em que o sistema de transporte aproxima-se do limite da sua capacidade.”*  
(Goodwin (2004) - tradução nossa)

Com essa definição, abstraímos dois principais fatores para caracterizar os congestionamentos: (1) a quantidade de veículos excedendo a capacidade da via, seja pelo excesso de veículos - como em horários de pico - ou pela diminuição da capacidade - como em casos de acidentes que interrompem o uso de algumas faixas; e (2) a redução da velocidade, total ou parcial. Usaremos também a expressão “**lentidão**” como sinônimo de “congestionamento”, em confluência com a nomenclatura usada no conjunto de dados da CET.

Assim como para os acidentes, há diversos meios para caracterizar a intensidade dos congestionamentos. Calatayud *et al.* (2021) detalha algumas medidas adotadas ao longo de quase um século de pesquisa, como a quantidade de veículos, tempo perdido no trajeto, e algumas mais sofisticadas como o “índice de atraso agregado”, que soma os custos de todos os veículos em um dado congestionamento. A escolha da medida depende da granularidade e do tipo de dados disponíveis. Neste trabalho, dimensionamos os trechos de lentidão pela sua **duração** (minutos) e pelo **comprimento do trecho** (em metros).

## 2.2 Inteligência Artificial

Para atingirmos os objetivos deste estudo, utilizaremos técnicas de Inteligência Artificial (IA). Uma definição clássica de IA, por Russell e Norvig (2009), indica que “IA é o estudo de agentes que recebem percepções do ambiente e executam ações” (tradução nossa). No nosso caso, o agente (modelo) deverá receber dados de congestionamentos e acidentes e executar ações para extrair conhecimento desses dados.

Para isso, transitaremos por duas sub-áreas da IA: Agrupamento de Dados e Regressão. Ambas as sub-áreas fazem parte do contexto mais amplo de **aprendizado de máquina**, que envolve o desenvolvimento de algoritmos e modelos que permitem que sistemas computacionais aprendam a partir de dados, sem serem explicitamente programados para realizar uma tarefa específica. O aprendizado de máquina é utilizado frequentemente em situações em que as regras para solução de um problema são desconhecidas ou muito complexas.

### 2.2.1 Agrupamento de Dados

Separar os dados em grupos (*clusters*) é uma tarefa importante na atividade humana. Crianças aprendem a separar cachorros e gatos, alimentos comestíveis e não comestíveis, etc. Muitos campos de conhecimento exigem tarefas de classificação de objetos semelhantes e automatizar essa atividade com algoritmos tem sido um tema de pesquisa ativo nos últimos 60 anos (Kaufman; Rousseeuw, 1990).

As técnicas de agrupamento de dados, ou clusterização (adaptado do inglês *clustering*), visam organizar elementos em grupos, baseando-se nas suas características. A ideia básica consiste em colocar em um mesmo grupo objetos que sejam similares de acordo com algum critério pré-determinado. Ou seja, os elementos de um determinado conjunto (*cluster*) devem ser “mutuamente similares e, preferencialmente, muito diferentes dos elementos de outros conjuntos” (Linden, 2009). Em geral, o agrupamento de dados é uma tarefa “não supervisionada”, ou seja, não depende de anotações humanas para treinar os algoritmos (Aggarwal, 2015). A Figura 1 mostra um exemplo de agrupamento de dados “similares”, em que os registros são agrupados por características comuns (no caso, a espécie). No contexto desta pesquisa, tais técnicas serão relevantes para agrupar os registros referentes a um mesmo evento de trânsito.

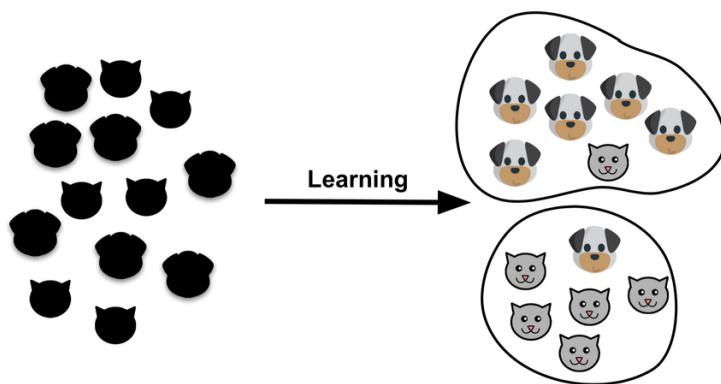


Figura 1 – Exemplo de Agrupamento de Dados. Fonte: Goyal (2018)

Há algumas abordagens (algoritmos) diferentes para executar a tarefa de agrupamento. O **Algoritmo K-Médias** (ou *K-Means*) exige que a quantidade de clusters (*K*) que deseja-se obter seja determinada antes do processamento. No início da execução, são

escolhidos aleatoriamente K pontos para serem os “centroïdes” dos clusters. Então, cada elemento é inserido no cluster cujo centroïde é o mais próximo. Depois, recalcula-se o centroïde dos clusters escolhendo o elemento mais ao centro ou extraíndo a média dos elementos. O processo de atribuição dos clusters aos elementos e redefinição dos centroïdes é repetido até que não seja mais possível diminuir mais a distância dos elementos ao centro dos seus clusters (Nielsen, 2016).

O algoritmo K-Médias é conhecido por sua eficiência no processamento, com rápida convergência dos clusters. No entanto, a necessidade de definir previamente a quantidade de clusters (K) pode representar um desafio, especialmente quando se trabalha com conjuntos de dados desconhecidos. Para contornar essa limitação, é possível testar diferentes valores de K e selecionar aquele que apresenta os melhores resultados, conforme as métricas de avaliação discutidas na subseção 2.2.3. Outra característica do K-Médias é a forma esférica dos clusters resultantes. Isso ocorre devido ao método de definição dos grupos, que se baseia na minimização da distância entre os elementos e os centros dos clusters.

O **Agrupamento Hierárquico**, por sua vez, é um processo iterativo que começa com clusters unitários (onde cada elemento constitui um cluster) e, a cada etapa, une os clusters mais próximos até que todos os elementos formem um único cluster. Existem várias alternativas para implementar essa fusão de grupos. Uma dessas abordagens é o *single link*, que define a distância entre dois clusters como a menor distância entre os elementos de ambos. Outra possibilidade é utilizar a distância média (*average link*) ou a distância máxima (*complete link*) entre os elementos dos clusters comparados. Após todas as fusões, gera-se uma árvore, chamada dendrograma, que contém todos os níveis de agrupamento, desde clusters unitários até um único cluster contendo todo o conjunto de dados, permitindo a escolha da quantidade de clusters mais adequada ao contexto dos dados (Nielsen, 2016).

Segundo Jarman (2020), o Agrupamento Hierárquico pode gerar resultados bastante distintos dependendo do método de fusão adotado. O uso da distância mínima entre clusters tende a formar clusters mais esparsos e populosos, pois elementos distantes podem ser agrupados se houver apenas um par de elementos próximos em seus grupos. Analogamente, o uso da distância máxima favorece a formação de clusters menores. Outro fator que influencia o resultado é a definição do “corte” na árvore resultante, ou seja, o ponto em que se decide interromper o processo de fusão. Esse corte pode ser feito ao definir a quantidade desejada de clusters ou ao estabelecer um valor limite de distância, de forma que elementos cujos grupos excedam essa distância sejam mantidos separados no resultado final.

Outra abordagem é o **Agrupamento por Densidade**, que identifica regiões densamente povoadas por elementos, com cada uma dessas regiões sendo considerada um cluster. O algoritmo DBSCAN processa o conjunto de dados percorrendo todos os pontos e, para cada ponto, encontra os elementos que estão acessíveis a uma distância mínima

pré-definida. Pontos que não são acessíveis a partir de outros elementos do conjunto são classificados como ruído. Os clusters resultantes podem assumir formas arbitrárias, desde que não haja interrupções na distribuição dos elementos (Ester *et al.*, 1996). Essa estratégia não requer a definição prévia da quantidade de clusters, mas possui um custo computacional mais elevado, pois, no limite, pode ser necessário calcular a distância entre todos os pares de pontos do conjunto. A Figura (2) compara os resultados das três abordagens descritas anteriormente, com os clusters representados por cores diferentes.

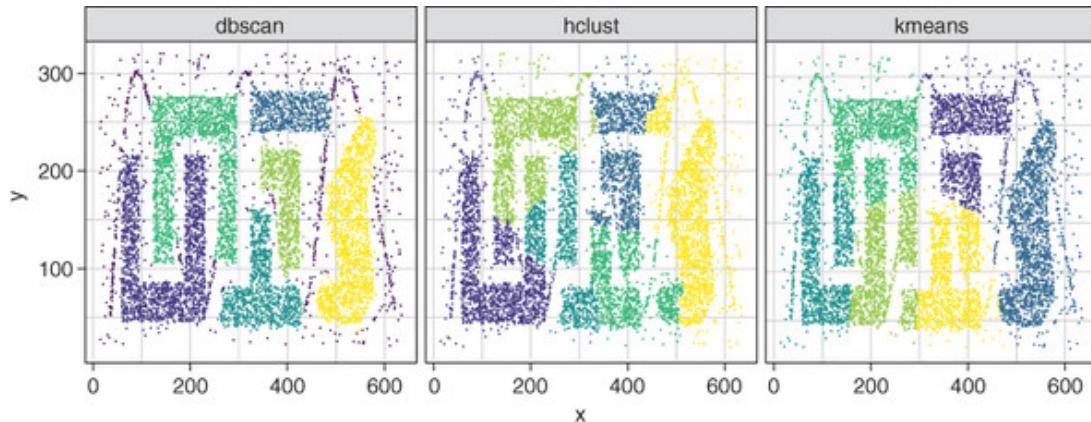


Figura 2 – Comparação entre métodos de agrupamento. Da esquerda para direita: agrupamento por densidade, hierárquico e K-médias. Fonte: Rhys (2020)

Para que os algoritmos de clusterização sejam eficazes, é crucial entender o conceito de “**distância**” (ou, inversamente, “**similaridade**”). Para cada tipo de objeto que deseja-se agrupar, deve-se definir medidas para calcular quão diferentes ou similares os elementos são uns dos outros. Considerando dados sobre acidentes de trânsito, poderíamos definir uma medida de similaridade que considera a quantidade de vítimas e valores de danos materiais, de modo que acidentes mais graves sejam agrupados juntos. Por outro lado, se desejarmos encontrar padrões nos momentos que os acidentes ocorrem, deveríamos agrupá-los por seus dias e horários, resultando em um agrupamento diferente. Os dados de congestionamentos e acidentes são exemplos de **dados complexos**, isto é, não podem ser comparados diretamente (como números), exigindo outras funções para extrair características e compará-los (Barioni, 2006). Detalharemos as métricas de distância utilizadas no Capítulo 3.

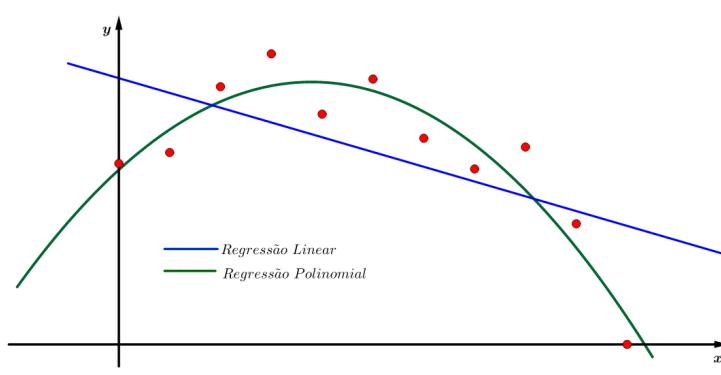
### 2.2.2 Regressão

Em aprendizado de máquina, um modelo de regressão busca resolver o problema de estimar os valores de uma função desconhecida. A tarefa de um modelo computacional de regressão é criar uma função que preveja o resultado dessa função desconhecida, dado um valor específico de entrada. Para isso, utiliza-se uma grande amostra de dados de entrada e saída, e o modelo de regressão determina as regras que serão aplicadas aos dados de entrada para gerar valores de saída que se aproximem dos observados. Em

outras palavras, o objetivo é determinar a relação entre uma variável de saída  $Y$  (também chamada de dependente ou resposta) e um conjunto de variáveis de entrada (conhecidas como independentes)  $(x_1, \dots, x_n)$  (Izbicki; Santos, 2020).

Por exemplo, podemos considerar que as variáveis independentes representam características de uma via (como o número de faixas, e a quantidade de semáforos), enquanto a variável resposta  $Y$  é a quantidade de horas em que houve congestionamento nessa via ao longo do ano. Ao aplicar um modelo de regressão, buscamos identificar as regras que determinam a influência de cada característica da via no aumento do congestionamento. O processamento desse exemplo poderia revelar que cada semáforo presente em uma via aumenta em 3,5 vezes a quantidade de horas anuais de congestionamento. Esse “peso” atribuído à variável “quantidade de semáforos” seria parte de uma modelagem estatística que deve ser avaliada por métricas específicas para determinar sua acurácia. Se essas métricas forem satisfatórias, teríamos identificado regras simples que simulam com precisão o comportamento de um problema complexo.

As técnicas de regressão podem ser classificadas de acordo com as funções geradas. Destacamos a **Regressão Polinomial**, que gera funções em formatos de polinômios  $\sum_{i=0}^n c_i x^i$ , em que  $n \in \mathbb{Z}$  é o grau máximo do polinômio. Na regressão polinomial, os valores de entrada são multiplicados por coeficientes (pesos) para gerar os valores de saída. Quanto maior o valor do grau  $n$ , conseguimos gerar funções que se adéquem melhor ao conjunto de dados original (também chamado de conjunto de treinamento). Entretanto, nesses cenários, além de aumentar a complexidade do modelo, aumenta-se o risco de **overfitting**, ou seja, o baixo poder de generalização para além dos dados de treinamento (Izbicki; Santos, 2020). Quando  $n = 1$ , denominamos o modelo de **Regressão Linear**, pois a função gerada representará uma reta (em duas dimensões) ou, genericamente, um *hiperplano* (Montgomery Elizabeth A. Peck, 2013). A figura 3 mostra um exemplo de regressão linear e polinomial, modelando uma função desconhecida com base em dados conhecidos (em vermelho).



(Figura 1)

Figura 3 – Regressão linear e polinomial. Fonte: UFRN (2018)

Uma técnica comum para implementar uma Regressão, que utilizaremos neste trabalho, é o **Método dos Mínimos Quadrados** (MMQ). Esse método obtém os parâmetros (coeficientes) minimizando a distância entre os valores reais da variável Y - originalmente conhecidos - e os valores preditos pela função  $f$  quando aplicada no conjunto de treinamento (Montgomery Elizabeth A. Peck, 2013).

Neste trabalho, aplicaremos a regressão para modelar a relevância que as características dos congestionamentos exercem na ocorrência de acidentes. Nossas variáveis independentes estarão relacionadas a lentidão das vias, e a variável dependente caracterizará os acidentes.

### 2.2.3 Métricas de Avaliação

Como determinar o algoritmo de agrupamento mais adequado? Como mensurar a qualidade das previsões de um modelo de regressão? Uma das abordagens mais comuns para a avaliação de modelos em aprendizado de máquina é a inspeção especializada dos resultados, que inclui técnicas de visualização e o uso de conhecimento prévio sobre as características do problema e dos algoritmos. No caso de algoritmos de agrupamento, a visualização de gráficos coloridos facilita a interpretação e avaliação da configuração dos clusters gerados. Para modelos de regressão, gráficos como o apresentado na Figura 3 ilustram a aderência do modelo aos dados observados.

Entretanto, a avaliação visual apresenta limitações. Além da subjetividade inerente ao julgamento humano, a análise de dados complexos - com muitas dimensões - torna inviável o uso exclusivo de técnicas de visualização. Ademais, grandes volumes de dados dificultam a análise manual, dada a sua demanda por tempo e recursos. Nesse contexto, o campo da Inteligência Artificial, e mais especificamente o aprendizado de máquina, requer o uso de métricas quantitativas e escaláveis para uma avaliação rigorosa e eficiente dos modelos.

#### 2.2.3.1 Métricas de Avaliação para Agrupamentos

Para os algoritmos de agrupamento de dados, há mais de 40 índices estabelecidos na literatura e ainda é um tema ativo de pesquisa (Hassan *et al.*, 2024). Neste trabalho, usaremos duas técnicas populares de avaliação de agrupamento. O **Coeficiente de Silhueta**, proposto por Rousseeuw (1987), é uma métrica clássica para avaliar a coesão e separabilidade dos clusters. Para cada ponto  $x$  pertencente a um cluster  $C$ , o valor  $a(x)$  mede a proximidade desse ponto em relação aos demais integrantes de seu cluster. Calculamos  $a(x)$  como a média das distâncias entre  $x$  e todos os outros pontos de  $C$ . Em contraste, a métrica  $b(x)$  quantifica a separabilidade entre os clusters e consiste na média das distâncias entre  $x$  e os elementos do cluster mais próximo a  $C$ . O coeficiente de

silhueta, por fim, é a medida

$$s(x) = \frac{b(x) - a(x)}{\max(b(x), a(x))}$$

Como exemplo, se o objeto  $x$  estiver isolado dentro do seu cluster, então  $a(x)$  será elevado, resultando em um coeficiente  $s(x)$  reduzido. Da mesma forma, caso  $x$  estiver muito perto de outro cluster, a distância  $b(x)$  se aproximará de 0, assim como o coeficiente  $s(x)$ . Entretanto, se  $x$  estiver distante de membros de outros grupos e próximo dos membros do seu grupo, então o coeficiente de silhueta apresentará um valor alto (próximo de 1) - indicando maior qualidade do agrupamento. O índice de Silhoueta do agrupamento é a média dos coeficientes  $s(x)$  para todos os  $x$  do conjunto de dados.

A métrica **DBCV** (*Density Based Clustering Validation*) é uma alternativa à métrica da Silhueta. O DBCV foi desenvolvido por Moulavi *et al.* (2014) para avaliar agrupamentos baseados em densidade, que podem assumir formas arbitrárias. Em situações geográficas, como rios, ruas ou linhas de transmissão, os agrupamentos costumam apresentar formas alongadas, e pontos de um mesmo grupo podem estar distantes, como a nascente e a foz de um rio. Nesse contexto, o coeficiente de silhueta penalizaria a qualidade do agrupamento, pois considera grupos não convexos como menos coesos.

A validação através do DBCV evita essa penalização ao considerar a distância máxima entre dois pontos de um agrupamento, em vez da média das distâncias entre todos os pontos. O cálculo desse índice é realizado através de algoritmos especializados em grafos, estruturas de dados que representam redes (objetos conectados) (Feofiloff; Yoshiko; Kohayakawa, 2011). A Figura 4 (em inglês) compara os dois métodos de validação de agrupamentos descritos anteriormente, destacando como o índice de Silhueta favorece agrupamentos esféricos, como aqueles resultantes do algoritmo K-Médias.

#### 2.2.3.2 Métricas de Avaliação para Regressão

Ao contrário dos métodos de agrupamento, que caracterizam-se por serem não supervisionados, os algoritmos de regressão contam com a supervisão de dados previamente rotulados com respostas corretas, ou seja, são supervisionados. Assim, a validação dos modelos baseia-se em técnicas estatísticas de comparação entre resultados preditos e as respostas reais dos conjuntos de entrada.

Para otimizar a capacidade de generalização, deve-se realizar essa comparação em amostras dos dados que não foram utilizados para o treinamento, os chamados conjuntos de testes. Caso a validação ocorresse em dados utilizados para treinar o modelo, haveria o risco de viés nos resultados, gerando métricas aparentemente satisfatórias, mas com baixo desempenho quando aplicadas a dados novos, diferentes daqueles do treinamento.

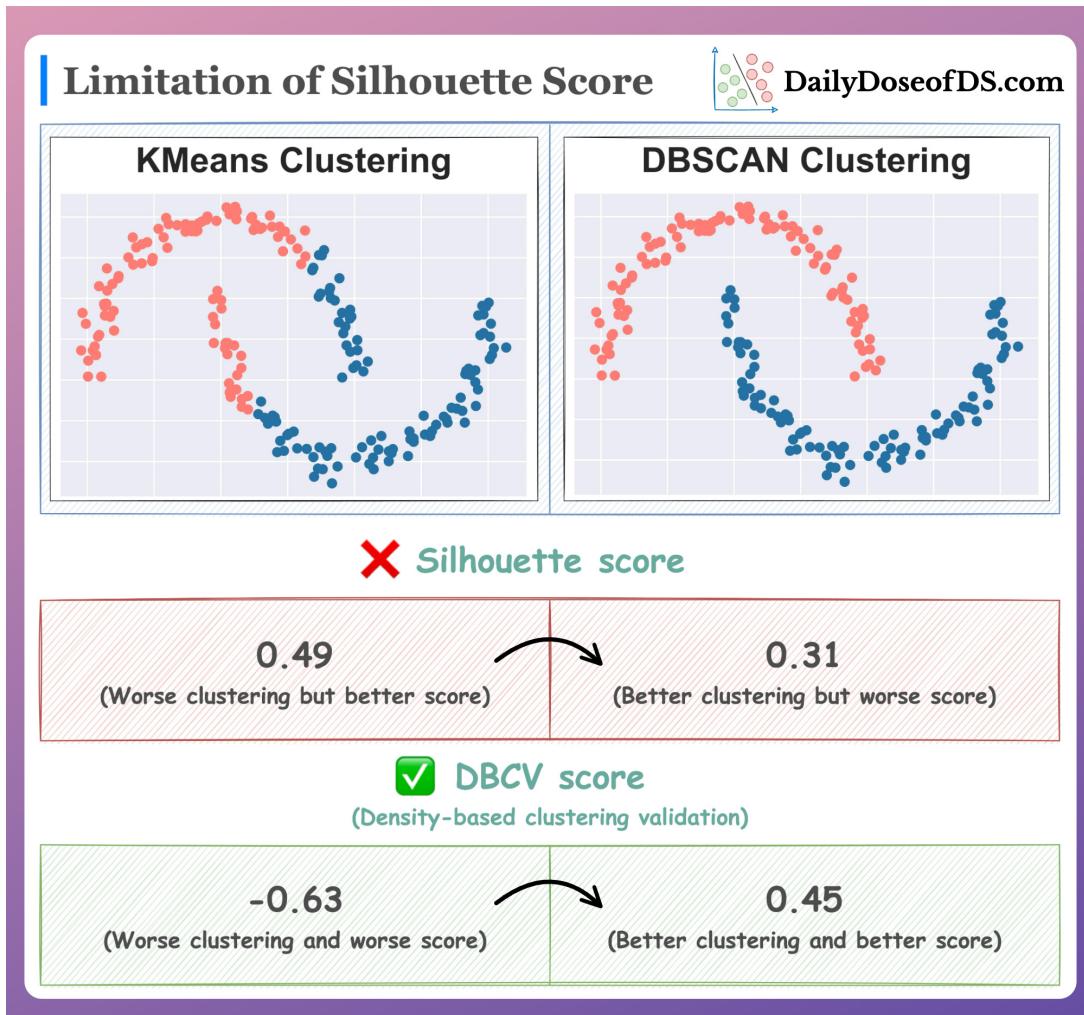


Figura 4 – Limitação do Coeficiente de Silhueta. Fonte: Chawla (2023)

Uma métrica popular de comparação entre os valores preditos e os observados é o **Erro Quadrático Médio**, ou MSE (*Mean Squared Error*). O MSE é definido pela expressão

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

onde  $n$  é a quantidade de dados do conjunto de testes,  $y_i$  são valores das variáveis respostas observados e  $f(x_i)$  é o valor predito pelo modelo, ou seja, o resultado da função gerada pelo treinamento, quando aplicada na  $i$ -ésima observação das variáveis independentes. Quando o MSE aproxima-se de zero, indica que o modelo produzido está adequado os dados observados (James *et al.*, 2021).

A segunda métrica que usaremos neste trabalho é o **Coeficiente de Determinação ( $R^2$ )**. Ele mede a proporção da variabilidade dos dados que é explicada pelo modelo. Pode ser interpretada como um indicativo da força da correlação entre as variáveis independentes e a variável dependente (Izbicki; Santos, 2020). Por exemplo, um  $R^2$  de 80% indica que o modelo de regressão é capaz de explicar grande parte do comportamento da variável dependente, restando apenas 20% da variabilidade que não é capturada pelas variáveis

independentes. Quando  $R^2$  é próximo de zero, significa que o modelo não consegue prever com precisão o comportamento da variável dependente. Para ilustrar, se tentássemos prever a altura de ondas de uma praia do Canadá utilizando como variáveis independentes a quantidade de vitórias de um time argentino de futebol, provavelmente obteríamos um  $R^2$  baixo, devido a ausência de correlação entre as variáveis.

O coeficiente  $R^2$  é calculado pela seguinte fórmula:

$$R^2 = 1 - \frac{\sum (y_i - f(x_i))^2}{\sum (y_i - \bar{y})^2}$$

, onde  $\bar{y}$  é média dos valores observados. Caso o modelo consiga predizer os valores de  $y$  com acurácia, então a soma do numerador se aproximará de 0, causando com que o valor do coeficiente seja próximo de 1 (Izbicki; Santos, 2020). Com essas métricas, conseguimos comparar diferentes implementações e extrair o conhecimento acurado de cada modelo.

### 2.3 Trabalhos Relacionados

A produção científica relacionada aos acidentes de trânsito tem se intensificado desde a década de 1990. Diversos estudos apontam a crescente relevância do problema, especialmente no que se refere aos impactos na vida das pessoas e na economia das sociedades. Mohammed *et al.* (2019) realizaram uma revisão de literatura com dados provenientes de diversos países, destacando os impactos “catastróficos” da falta de segurança no trânsito, que ocasiona prejuízos sociais e econômicos até oito vezes maiores que os decorrentes de conflitos armados em países do Oriente Médio.

A análise das causas dos acidentes de trânsito tem sido um campo de pesquisa ativo nas últimas três décadas (Chaudhary; Bhaduria; Garg, 2017). A descoberta de fatores que influenciam significativamente a ocorrência de acidentes, como a velocidade dos veículos e consumo de álcool (Chand; Jayesh; Bhasi, 2021), permitiu que autoridades implementem ações para reduzir o volume dos acidentes. No entanto, os países em desenvolvimento ainda carecem de soluções eficazes para a redução desses índices. Apesar de possuírem 60% da frota de veículos, esses países são responsáveis por 92% dos acidentes de trânsito no mundo (WHO, 2023). Nessas regiões, desafios estruturais e políticos impõem a necessidade de estudos especializados para enfrentar o problema.

No Brasil, até os anos 2000, a pesquisa sobre acidentes de trânsito ainda era pouca desenvolvida (Vasconcellos, 1999). Um dos principais desafios nessa área é a obtenção de dados, que muitas vezes não são registrados de forma completa ou com o nível de detalhamento apropriado (Bacchieri; Barros, 2011). Recentemente, relatórios analíticos, como o desenvolvido pela CET-SP (2022), tem realizado análises descritivas do perfil das ocorrências, destacando, por exemplo, a alta participação de motocicletas nos acidentes e a utilização de mapas que ilustram as regiões com maior incidência de sinistros.

Ao longo dos anos, o estado da arte de diversas sub-áreas da mobilidade urbana passou a se basear em modelos de Inteligencia Artificial aplicados a grandes volumes de dados (Almukhalfi; Noor; Noor, 2024). No contexto de Big Data, grandes quantidades de dados são geradas constantemente por sensores e dispositivos conectados à internet, a chamada *IoT - Internet of Things* ou Internet das Coisas). Assim, houve uma transição dos modelos tradicionais - em que os dados são coletados e analisados manualmente por especialistas - para modelos que utilizam conjuntos massivos de dados, exigindo processos computacionais mais sofisticados. De acordo com Almukhalfi, Noor e Noor (2024), tópicos como predição de congestionamentos, predição de acidentes, determinação de tempos de semáforos, entre outros, destacam-se cada vez mais pelo uso de técnicas de IA.

No âmbito dos acidentes de trânsito, Zohra *et al.* (2023) analisaram o desempenho de algumas técnicas de predição, alcançando uma acurácia de até 90% na previsão das ocorrências de acidentes com base em dados do ambiente e da localização. Os autores analisaram um conjunto de dados de sinistros nos Estados Unidos e aplicaram três modelos de aprendizado de máquina. Os resultados destacaram que fatores como a topografia da via, especialmente a configuração de cruzamentos, acessos e semáforos, influenciam significativamente a gravidade dos acidentes.

Estudos também têm explorado a relação entre os congestionamentos e os sinistros de trânsito, como os realizados por Retallack e Ostendorf (2019), encontraram diversos padrões de correlação, incluindo correlações positivas, negativas, e em formato de U, em que os acidentes são mais intensos apenas quando há pouco ou muita lentidão. Outros estudos também concluíram que congestionamentos não afetam a ocorrência ou gravidade dos acidentes de trânsito (Quddus; Wang; Ison, 2010). Calatayud *et al.* (2021), por sua vez, utilizaram dados de aplicativos de trânsito de dezenas de cidades e identificaram forte correlação positiva, ou seja, quanto maior a extensão da lentidão, maior a probabilidade da ocorrência dos incidentes e sinistros. Li, Gui e Liu (2022) utilizaram uma versão do Algoritmo K-Médias para identificar os padrões dos congestionamentos de Pequim, concluindo que diferentes distritos possuem diferentes padrões de tráfego de acordo com a distância do centro da cidade.

Em suma, este trabalho baseia-se em técnicas amplamente utilizadas pela literatura existente do tema estudado. Almejamos aplicar essas técnicas nos conjuntos de dados do estudo de caso de São Paulo para produzir conhecimento específico sobre os padrões da cidade e extrapolar as conclusões para a ainda ativa discussão científica sobre o tema.



### 3 METODOLOGIA

Diversas técnicas de diferentes sub-áreas da computação foram propostas neste trabalho para atingir o objetivo de identificar a correlação entre congestionamentos e acidentes de trânsito. A Figura 5 ilustra as principais etapas do desenvolvimento, que serão detalhadas ao longo deste capítulo.

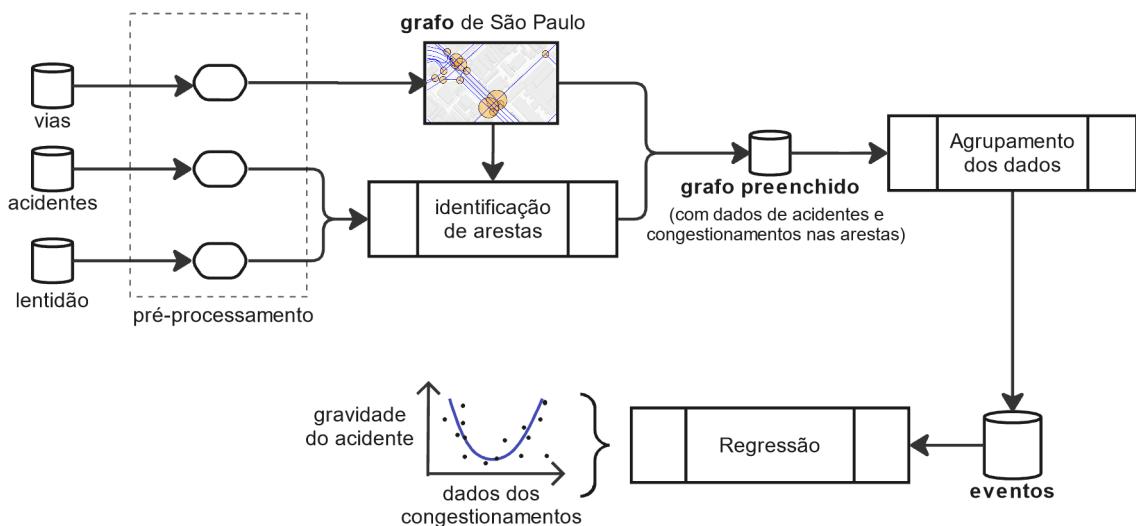


Figura 5 – Metodologia do trabalho

As primeiras atividades, na linha superior da figura, (pré-processamento, identificação de arestas e agrupamento dos dados) têm o objetivo de integrar os dois datasets estudados (acidentes e congestionamentos) em uma estrutura de dados adequada. Cada um deles tem características próprias para descrever e localizar os registros que precisam ser compatibilizadas. Ao final dessas etapas, geramos uma base de “eventos”, que são objetos com características (*features*) de congestionamentos, acidentes e das vias da cidade. Posteriormente, o dataset de eventos é a entrada (*input*) dos algoritmos de regressão que identificam os padrões que propusemos examinar. A seguir, detalhamos as implementações técnicas de tais atividades.

#### 3.1 Pré-Processamento

As etapas de pré-processamento são responsáveis pela transformação e limpeza dos dados brutos, garantindo que estejam em um formato adequado para que os conjuntos de dados possam ser integrados e correlacionados.

### 3.1.1 Vias

Antes de lidar com os dados de congestionamentos e acidentes, é necessário construir o arcabouço de todo o processamento: o grafo da cidade de São Paulo.

No contexto geográfico de uma cidade, construiremos um grafo em que as arestas são os logradouros (ruas, avenidas, pontes) e os vértices são os pontos em que os logradouros terminam ou se conectam (cruzamentos, esquinas). A Figura 6 ilustra o grafo construído para uma região da cidade de São Paulo, com o mapa ao fundo. Os vértices estão representados por círculos laranjas e as arestas por linhas azuis. Quanto maior o raio do círculo, mais ligações estão conectadas, ou seja, maior o grau do vértice.

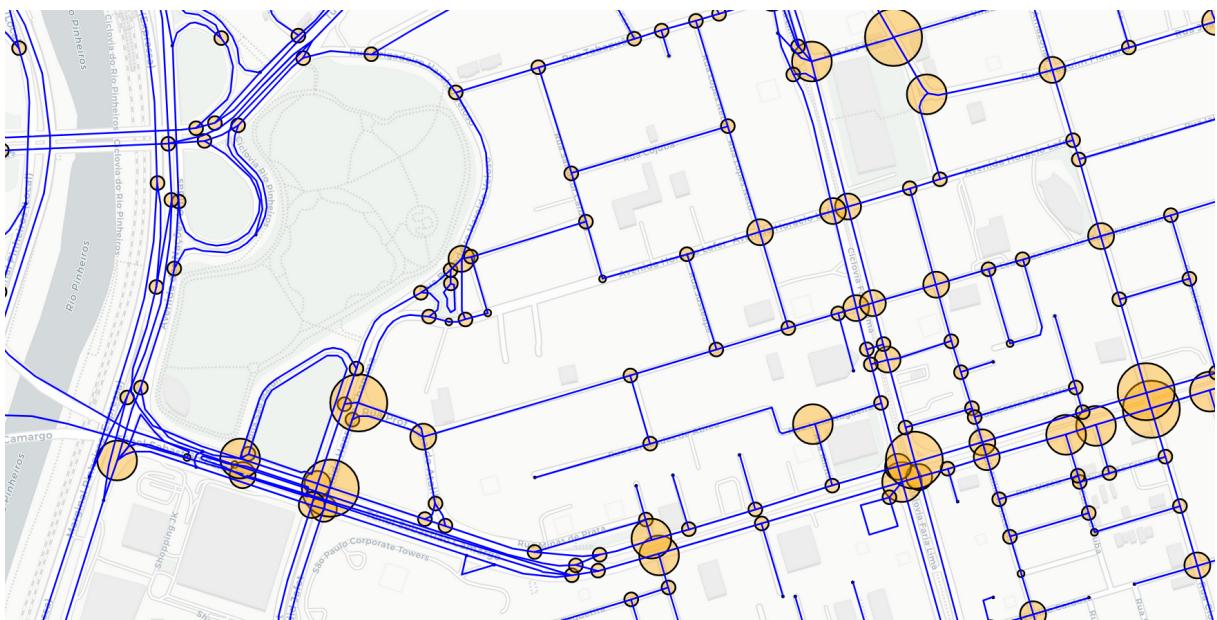


Figura 6 – Grafo da cidade de São Paulo

O conjunto de dados base para a confecção do grafo é o dataset de logradouros da prefeitura de São Paulo, disponibilizado no sistema *GeoSampa*<sup>1</sup>. Nesse dataset, há uma linha para cada trecho de via. Ou seja, em uma avenida com cruzamentos, o trecho entre cada par de ruas (um quarteirão) será um registro diferente. Cada registro contém o nome do logradouro e a sequência de coordenadas (latitude e longitude) do trecho.

O algoritmo para construção do grafo possui três etapas. Primeiramente, geramos os conjuntos de vértices e arestas a partir das linhas do dataset. Para isso, cada linha origina dois vértices (o ponto inicial e o ponto final do trecho) e uma aresta (a via em si). O grafo gerado contém 185 mil vértices e 222 mil arestas.

A segunda etapa consiste em unir vértices que estão muito próximos - por conta de imprecisões numéricas. Tal tratamento é necessário devido à representação das coordenadas por pontos flutuantes, ocasionando que pontos que conectam dois trechos contínuos de uma

<sup>1</sup> [https://geosampa.prefeitura.sp.gov.br/PaginasPublicas/\\_SBC.aspx](https://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx)

via nem sempre coincidam com exatidão. Como consequência, uma via torna-se desconexa, com diversos vértices muito próximos, mas não conectados. Então, utilizamos o algoritmo KDTree (SciPy, 2024) para identificar pontos próximos e uni-los, concatenando as arestas dos vértices originais. Antes dessa etapa, 79% das arestas estavam conectadas. Após o processamento, o maior componente conexo do grafo abrangia 97% da cidade e o número de vértices foi reduzido para 156 mil.

Por último, propomos uma redução de dimensionalidade do grafo - visando melhorar a eficiência dos algoritmos sem perder informações. Nessa etapa, identificamos vértices com apenas duas arestas incidentes - ou seja, aqueles que são apenas um ponto de “conexão” entre dois outros vértices. Nesses casos, podemos concatenar as duas arestas e eliminar o vértice original. Repetindo o processo iterativamente, diminuímos 13 mil vértices e 13 mil arestas. Após as 3 etapas de pré-processamento, conclui-se a geração do grafo que hospedará os demais dados deste estudo.

### 3.1.2 Acidentes

Os dados sobre acidentes de trânsito, obtidos do sistema InfoSiga, estão originalmente organizados em dois arquivos diferentes - acidentes fatais e acidentes não fatais. A estrutura dos arquivos são semelhantes. Ambos contém dados sobre o momento do acidente (data, hora e dia da semana); localização (endereço e coordenadas); tipo do sinistro (colisão, atropelamento, etc); e gravidade. Na base dos acidentes fatais, há a quantidade de vítimas e outros dados sobre o óbito.

O pré-processamento para os dados de acidentes possui quatro fases: (1) tratamento de duplicatas, (2) unificação dos datasets, (3) preenchimento de valores nulos e (4) filtragem pela localização.

Como temos arquivos distintos, a primeira etapa visa garantir a unicidade dos registros. Assim, criamos uma chave textual que é a composição do momento e localização (endereço) do acidente. Com esse método, identificamos 527 registros duplicados entre os dois arquivos - que tiveram as cópias desconsideradas. A chave também foi usada para investigar outros arquivos do sistema InfoSiga (como arquivos sobre acidentes em rodovias) em que conferimos que seus registros estavam contidos nos dois conjuntos de dados já contemplados. Posteriormente, essa chave foi mapeada para uma chave numérica, para simplificar os processamentos.

Após a criação de um identificador único para cada registro, pode-se unir os dois datasets. Para isso, padronizamos os nomes das colunas - que têm algumas diferenças em cada conjunto. Depois, o atributo de tipo de acidente é categorizado com base nas definições detalhadas na subseção 2.1.1: fatal, não fatal com vítimas (atropelamentos), não fatal sem vítimas (colisões) e incidentes (sem vítimas nem danos). Por fim, diferentes colunas de quantidade de vítimas são somadas em um só atributo.

O dataset resultante contém 959 mil registros. Nele, há registros cuja descrição da localização está incompleta - exigindo a terceira etapa do pré-processamento: tratamento de dados nulos. Os dados contém colunas para o endereço do acidente e outras duas colunas para as coordenadas (latitude e longitude). Para os casos em que há o endereço completo, mas não as coordenadas, utilizamos a ferramenta (API) gratuita e de código aberto *OpenStreetMap*<sup>2</sup> para obter as coordenadas do local. Há 12 mil registros da cidade de São Paulo nessa condição. Por outro lado, há registros sem dados de coordenadas nem de endereço. Esses casos representam 2,5% do total e serão excluídos do dataset para a continuidade do estudo - uma vez que não é possível associá-los aos congestionamentos.

A última etapa do pré-processamento é o filtro de acidentes com base em sua localização. O sistema InfoSiga registra dados de todo o estado de São Paulo. No entanto, os dados de congestionamentos (descritos na próxima seção) abrangem apenas grandes corredores de uma região da cidade de São Paulo, conhecida como “centro expandido”. Dessa forma, foram mantidos no dataset apenas os acidentes das regiões presentes nas bases de congestionamento, conforme ilustrado na Figura 7. Essa abordagem evita o risco de viés nas análises, pois a manutenção desses registros resultaria em regiões com sinistros, mas sem dados de congestionamento, o que poderia levar à falsa impressão de ausência de lentidão, quando se trata da ausência de dados. O dataset resultante do processamento descrito contém 82 mil registros, referentes aos anos de 2019 a 2023.

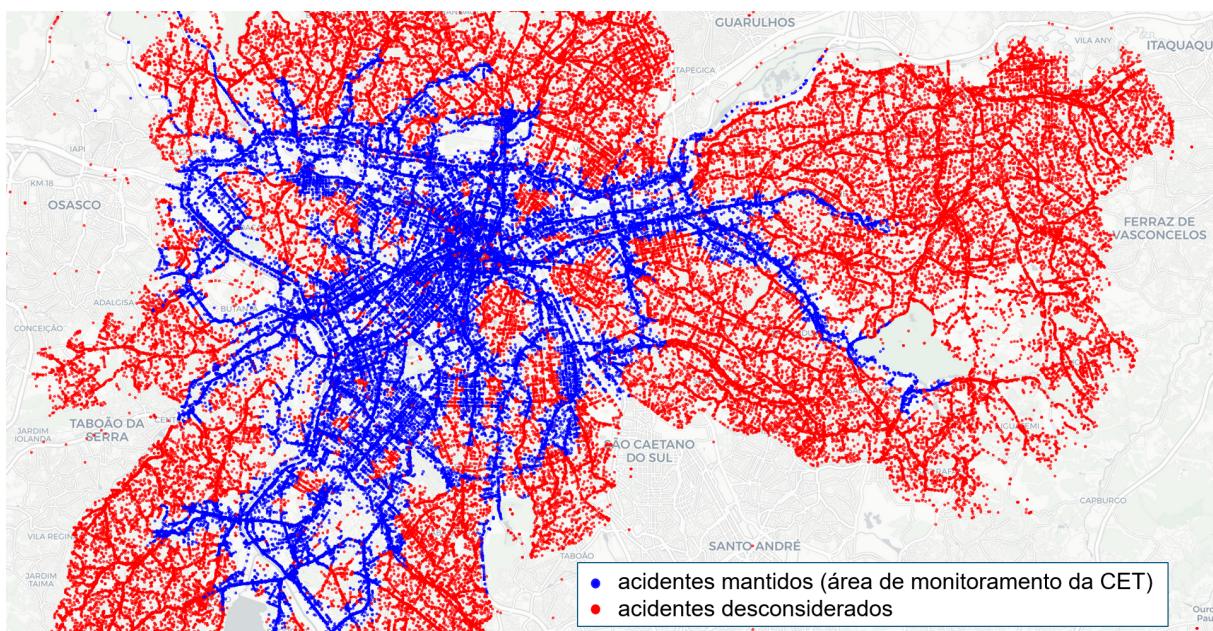


Figura 7 – Posicionamento dos acidentes de trânsito em relação à área de monitoramento dos congestionamentos

<sup>2</sup> <https://nominatim.openstreetmap.org/ui/>

### 3.1.3 Congestionamentos

Os conjuntos de dados de congestionamentos, fornecidos pela Companhia de Engenharia de Tráfego da cidade de São Paulo (CET-SP), estão disponibilizados separadamente por ano. Portanto, a primeira etapa do processamento consiste em integrar os arquivos de 2019 a 2023 - mesmo período dos dados disponíveis sobre acidentes.

O principal desafio dos dados sobre as lentidões é que os registros não são georreferenciados. Considere o registro a seguir, denominado  $C_1$ : “Congestionamento de 1567 metros, no corredor Marginal Pinheiros, de 593 m antes de ALEXANDRE MACKENZIE até 2030 m depois de CIDADE UNIVERSITARIA”. Note como a descrição é textual e não estruturada. Assim, deve-se desenvolver um processamento capaz de localizar as coordenadas geográficas representadas textualmente por cada registro.

A estratégia escolhida para esse desafio caracteriza-se por vincular as descrições dos trechos de lentidão com as arestas do grafo construído - que contém o nome dos logradouros. Essa etapa será descrita na próxima subseção. Contudo, para que essa computação seja eficaz, fez-se necessário alguns ajustes manuais nos nomes das vias presentes nos registros. Em suma, retiramos abreviaturas e trocamos nomes populares, como o viaduto “cebolinha”, pelos nomes oficial, como “VD JOAO JORGE SAAD”. Ajustamos manualmente 70 nomes de vias.

A base de congestionamentos contém 830 mil registros distribuídos em cerca de 200 conjuntos de logradouros (também chamados corredores). Essencialmente, apenas as grandes avenidas da cidade são abrangidas, sem ruas locais dentro dos bairros. Essa característica pode ser explicada pois a monitoração da CET ocorre de forma semi-automatizada, em que os agentes de trânsito registram o fluxo de veículos com o apoio de sensores e câmeras apenas nas localizações com maiores fluxos de veículos (CET-SP, 2024). A anotação manual dos agentes também explica a característica textual dos registros descrita anteriormente.

## 3.2 Identificação de Areias

O objetivo desta fase do trabalho é integrar os dados dos acidentes e dos congestionamentos na estrutura de grafo que montamos. Para o caso dos acidentes, conforme descrito na Subseção 3.1.2, temos as coordenadas geográficas de cada ocorrência. Dessa forma, pudemos aplicar o algoritmo KDTTree (o mesmo utilizado durante a construção do grafo) para localizar qual vértice é o mais próximo daquele acidente. Depois de obter o vértice, percorremos as arestas conectadas a ele para detectar qual é a mais próxima do local do acidente, através do cálculo da distância entre ponto e reta implementado pela biblioteca de código Shapely (2024). Ilustramos esse procedimento na Figura 8.

A base dos trechos de lentidão, por sua vez, exige um processamento mais sofisticado.

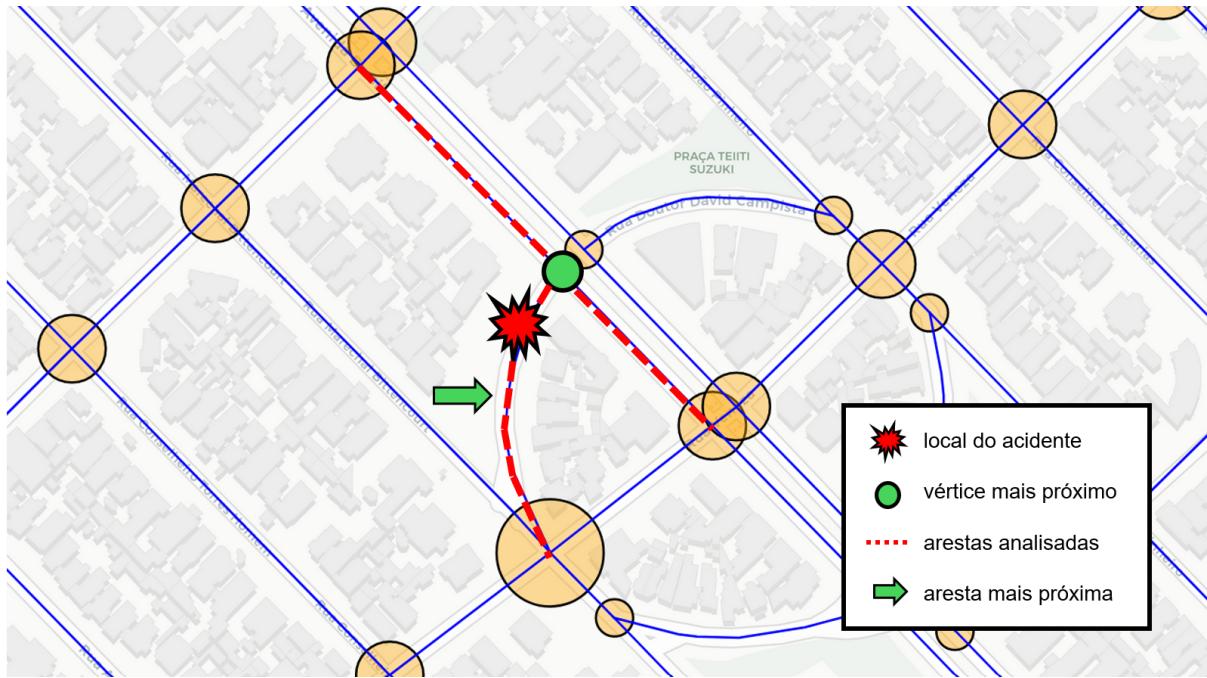


Figura 8 – Procedimento para identificação da aresta mais próxima a um acidente

Conforme descrito na Subseção 3.1.3, a abordagem escolhida consiste em um algoritmo para encontrar as arestas do grafo cujo nome possui maior similaridade com a descrição textual do congestionamento. Importamos a função de distância da biblioteca *difflib* (Python, 2023). Após o pré-processamento, foi possível definir as arestas correspondentes a todas os corredores de tráfego presentes na base de congestionamentos - conforme mostrado na Figura 9.

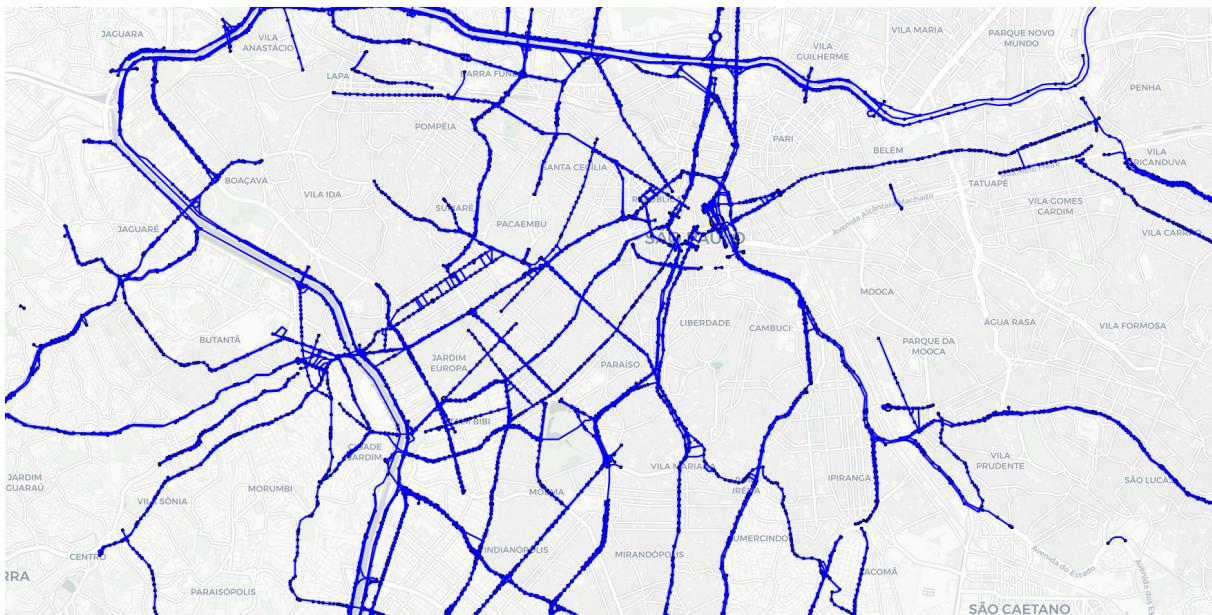


Figura 9 – Subgrafo da cidade correspondente às vias monitoradas pelo sistema de registro de congestionamento da CET-SP

Após identificar as arestas das vias, deve-se localizar qual trecho (ou seja, qual

subconjunto das arestas) correspondem a cada registro. Voltando ao exemplo  $C_1$  (detalhado em 3.1.3), não basta identificarmos as arestas do corredor Marginal Pinheiros. Devemos restringi-las àquelas entre a avenida Alexandre Mackenzie e a ponte Cidade Universitária.

O procedimento descrito a seguir está presente na figura 10. Primeiramente, obtemos as arestas da 3-vizinhança do subgrafo da via - isto é, consideramos as arestas da via, junto das arestas que estão conectadas nela (vizinhas), e também as arestas que estão conectadas nas vizinhas (vizinhas das vizinhas) e assim sucessivamente. Então, buscamos nesse novo conjunto de arestas os pontos de referência do início e do fim do trecho. No caso de  $C_1$ , encontramos uma aresta da avenida Alexandre Mackenzie e uma aresta da ponte Cidade Universitária, e traçamos o caminho mais curto entre elas. Por fim, fazemos uma reorientação do caminho para que a distância obtida seja compatível com a distância declarada na descrição - ou seja, deslocamos os pontos de início e término até que o comprimento das arestas se aproxime ao máximo do valor esperado.

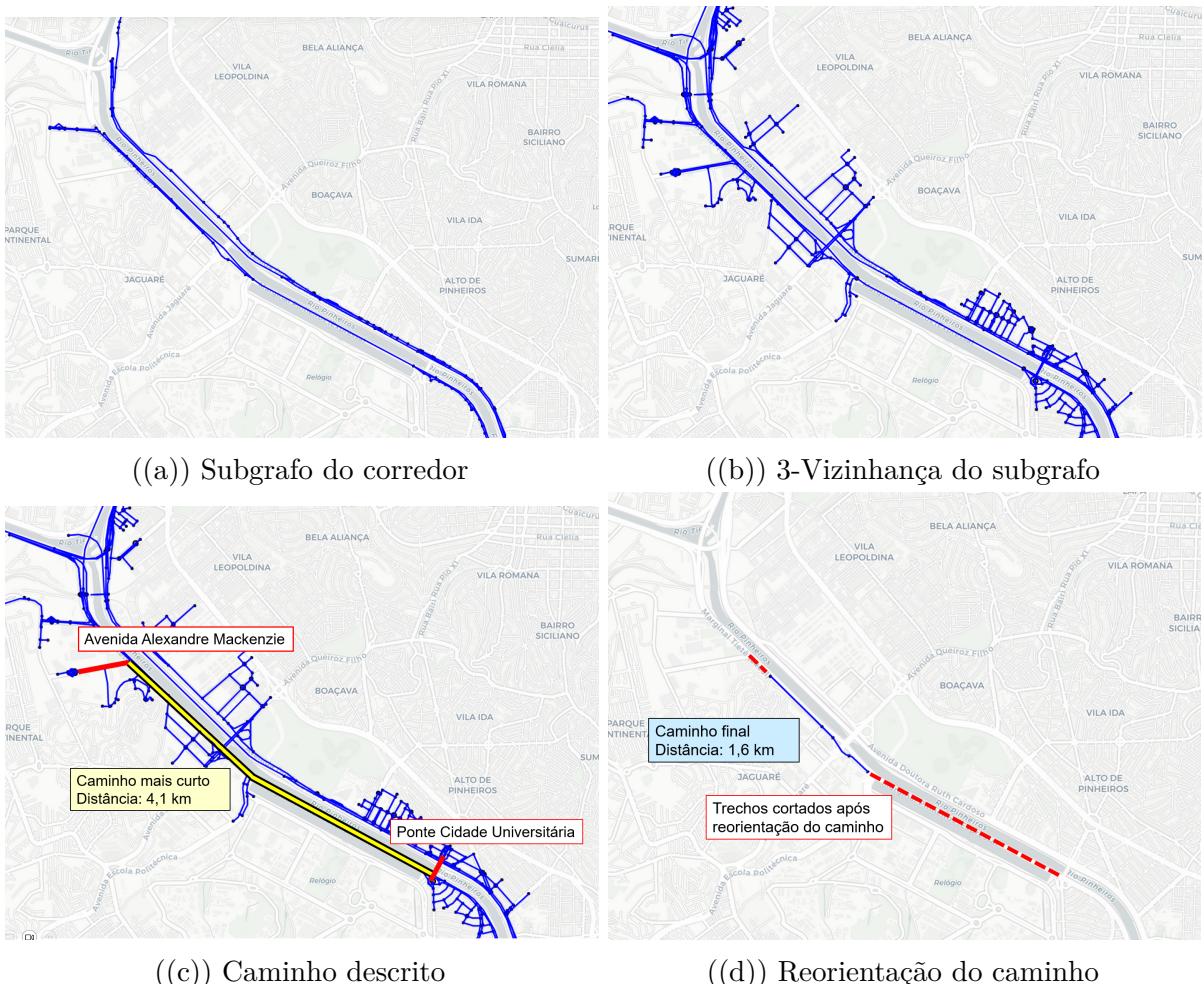


Figura 10 – Execução do algoritmo de identificação de arestas para o exemplo de congestionamento  $C_1$

Após as identificações, os registros de acidentes e congestionamentos são associados às arestas encontradas. Dessa forma, o grafo torna-se o repositório central de todos os

dados manipulados neste estudo. A estrutura do grafo é fundamental para a integração - de semântica geográfica - dos dois temas investigados neste trabalho. Ademais, o grafo destaca-se por sua capacidade de visualização dos dados, oferecendo uma representação direta na forma de mapas, como aqueles exibidos nessa seção.

### 3.3 Agrupamento dos dados

As etapas iniciais da metodologia, conforme descritas anteriormente, concentram-se na organização dos dados. Após a execução dessas etapas, um volume considerável de dados está distribuído ao longo das arestas do grafo. Contudo, apesar de estarem integrados em uma única estrutura, ainda não é possível estabelecer relações diretas entre eles. O agrupamento de dados representa, por sua vez, um primeiro passo para a extração de conhecimento e a identificação de padrões.

O objetivo deste processamento é unir arestas que correspondam a uma mesma localidade em um mesmo período, conforme introduzido na Subseção 2.2.1. Por exemplo, três registros de congestionamentos simultâneos em ruas que se cruzam se tratam na verdade de um mesmo fenômeno de lentidão. Da mesma forma, dois acidentes com 15 minutos de intervalo entre eles em uma mesma via congestionada devem ser interpretados como estando relacionados. Assim, há apenas duas dimensões que devem ser observadas para essa finalidade: o espaço e o tempo.

Antes de aplicar os algoritmos de clusterização, precisamos transportar os dados do grafo para um formato mais adequado. Esse novo formato será um dataset das arestas do grafo, porém, cada registro de congestionamento ou acidente vinculados a uma aresta ocasionará uma réplica dessa aresta. Por exemplo, se houve 50 observações de congestionamentos na aresta A ao longo dos anos e também 5 registros de acidentes, então haverá 55 linhas (também chamadas de “tuplas”) vinculadas a aresta A no dataset. Cada tupla possuirá 4 atributos: (1) identificação da aresta; (2) identificação do registro vinculado (acidente ou congestionamento); (3) data e hora do registro vinculado; (4) coordenadas da aresta<sup>3</sup>. Os dois primeiros atributos serão utilizados apenas para rastreio dos registros originais a fim de recuperar as demais informações posteriormente. O dataset de arestas possui 5 milhões de registros e será o input dos métodos de clusterização.

Em consonância com as definições do Referencial Teórico, determinamos um meio de comparar os elementos que serão agrupados, ou seja, definir uma função de distância. Para isso, consideramos o tempo como uma terceira dimensão, de modo que as ocorrências serão medidas pela latitude, longitude e momento. Os dados foram transformados em escala de modo que uma hora equivalesse a 100 metros de distância nas variáveis espaciais.

<sup>3</sup> Na biblioteca de código Shapely, utilizada para esse processamento, os dados geográficos são armazenados em um único atributo chamado *geometry*, independentemente da quantidade de coordenadas e do tipo de objeto geométrico.

Na prática, a função de distância produzirá o mesmo valor ao comparar dois acidentes que ocorreram no mesmo momento a uma distância de 500 metros ou no mesmo local mas com 5 horas de intervalo. Essa abordagem foi escolhida com avaliações empíricas das ocorrências, visando para balancear as duas dimensões.

Para o processamento dos algoritmos de agrupamento, dividimos o dataset de arestas por dias - já que os congestionamentos têm sazonalidades de algumas horas. Essa divisão resultou em ganhos significativos de eficiência e acurácia. Dessa forma, para cada dia dos 4 anos contemplados (2019 a 2023), executamos as três alternativas expostas na seção 2.2.1 (K-médias, Hierárquico e DBSCAN), sendo que cada execução exigiu a definição de parâmetros com melhor desempenho. Os resultados de cada abordagem serão detalhados no Capítulo 4.

### 3.3.1 Eventos

Os conjuntos de ocorrências decorrentes do agrupamento realizado determinam o objeto principal deste estudo: os **eventos**. Um evento é o tipo de registro que - finalmente - correlaciona os congestionamentos e os acidentes, sendo composto dos agrupamentos das ocorrências junto dos registros originais que as compuseram.

Especificamente, um evento possui os seguintes atributos:

- Arestas correspondentes (apenas para rastreio);
- Momento inicial;
- Momento final;
- Duração do evento (subtração do fim pelo inicio);
- Dia da semana;
- Tamanho máximo do congestionamento;
- Quantidade de vértices;
- Média do grau dos vértices;
- Quantidade de acidentes;
- Quantidade de vítimas;
- Gravidade do acidente (conforme 2.1.1).

Isso posto, o objeto do evento encapsula as dimensões temporal, geográfica (pela topologia dos vértices e arestas), dos congestionamentos (tamanho e duração) e dos sinistros de trânsito. Enquanto os atributos estruturados viabilizarão os modelos de regressão, a estrutura implícita do grafo permite a inspeção visual e a interpretabilidade dos resultados.

### 3.4 Regressão

A última etapa do processamento dos dados neste estudo envolveu a aplicação de uma análise de regressão (cuja teoria foi introduzida em 2.2.2) para quantificar a correlação entre os diferentes atributos dos eventos. A variável de interesse (dependente) é a gravidade do acidente. As variáveis independentes são os demais atributos dos eventos - com exceção dos momentos iniciais e finais e das arestas correspondentes - que já tiveram seus dados extraídos para os outros atributos.

Para aplicar os modelos de regressão, é necessário transformar a variável de interesse, que é categórica, para valores numéricos. Para isso, mapeamos cada categoria de gravidade do acidente para um número entre 0 e 1. A variável vale 0 se não houve acidentes e vale 1 se houve acidentes fatais com vítimas. Além do valor atribuído para a gravidade, há ainda um acréscimo referente à quantidade de vítimas e de acidentes: quanto mais vítimas naquele evento, maior será o valor da variável dependente. A figura 11 representa os valores convertidos.

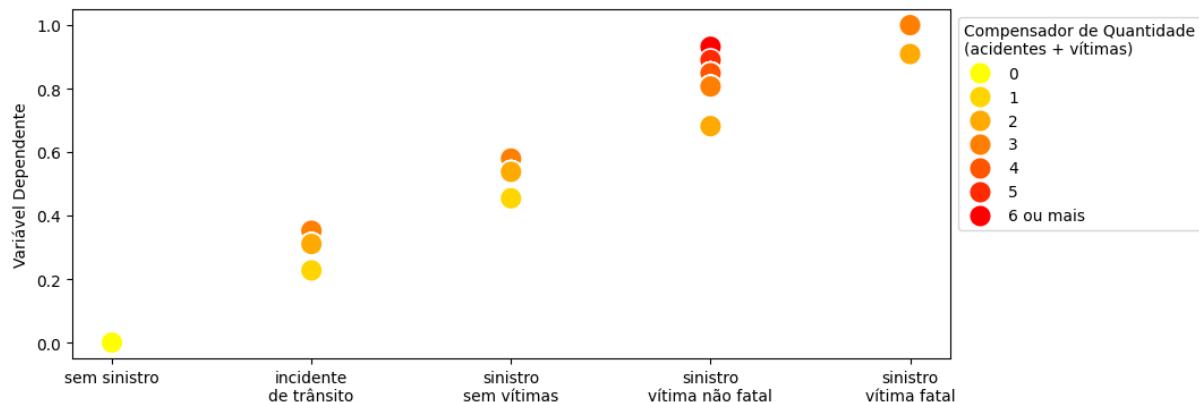


Figura 11 – Valores numéricos atribuídos à variável dependente (gravidade do acidente)

Antes da aplicação dos modelos, o dataset de eventos passou por uma padronização dos dados para que as variáveis numéricas estivessem em uma mesma escala. Utilizamos a estratégia de transformar as variáveis para que se distribuíssem ao redor do zero, com desvio padrão 1, isto é, seguissem uma distribuição Normal(0,1) (Scikit, 2024). Os conjuntos de testes foram separados aleatoriamente com um volume de 20% dos registros. A execução da regressão polinomial foi realizada com grau 1 (linear), 2, 3 e 4. Os resultados serão discutidos a seguir.

## 4 ANÁLISE DOS RESULTADOS

Neste capítulo, avaliaremos as diferentes implementações dos algoritmos de agrupamento e regressão. Discutiremos quais modelos obtiveram melhores desempenhos e quais conhecimentos pudemos extrair dos dados.

### 4.1 Algoritmos dos Agrupamentos

Antes de analisar os resultados das técnicas de clusterização com métricas quantitativas, inspecionamos visualmente os resultados. Vamos utilizar como exemplo o registro de acidentes e congestionamentos do dia 30/08/2019, com 4700 arestas, conforme ilustrado na Figura 12.

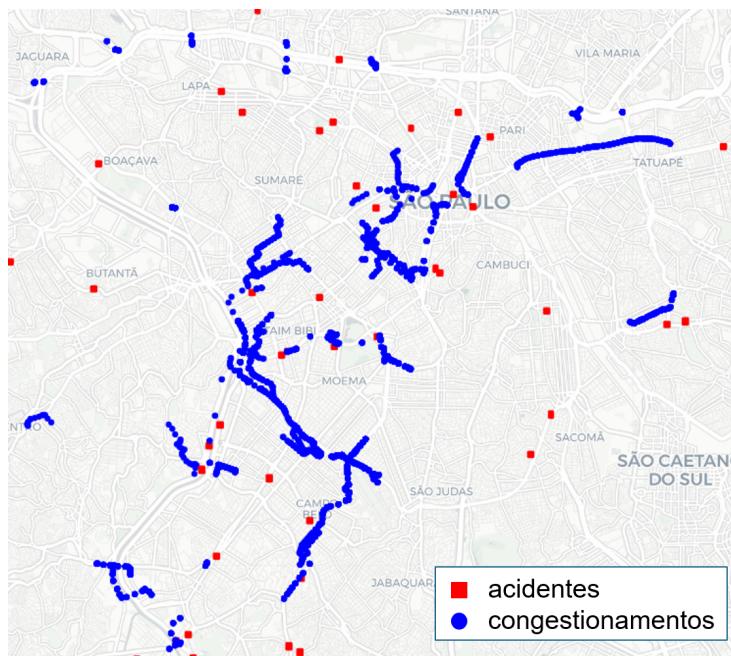


Figura 12 – Mapa de acidentes e congestionamentos do dia 30/08/2019 na cidade de São Paulo

Os clusters resultantes dos algoritmos K-médias, Hierárquico e DBSCAN são comparados na Figura 13. Observa-se que a estratégia de K-Médias apresenta limitações na identificação de clusters com formatos alongados, como ruas e avenidas, o que leva à divisão de um mesmo evento de congestionamento em vários grupos. O agrupamento hierárquico, por sua vez, conseguiu capturar clusters com formas alongadas, mas há interseções visíveis entre alguns clusters. Essa aparente interseção pode ser explicada pela dimensão temporal, que não está representada no mapa. O algoritmo DBSCAN produziu resultados visualmente mais coerentes, identificando de forma mais precisa eventos de congestionamento e acidentes ocorridos em proximidade geográfica e temporal.

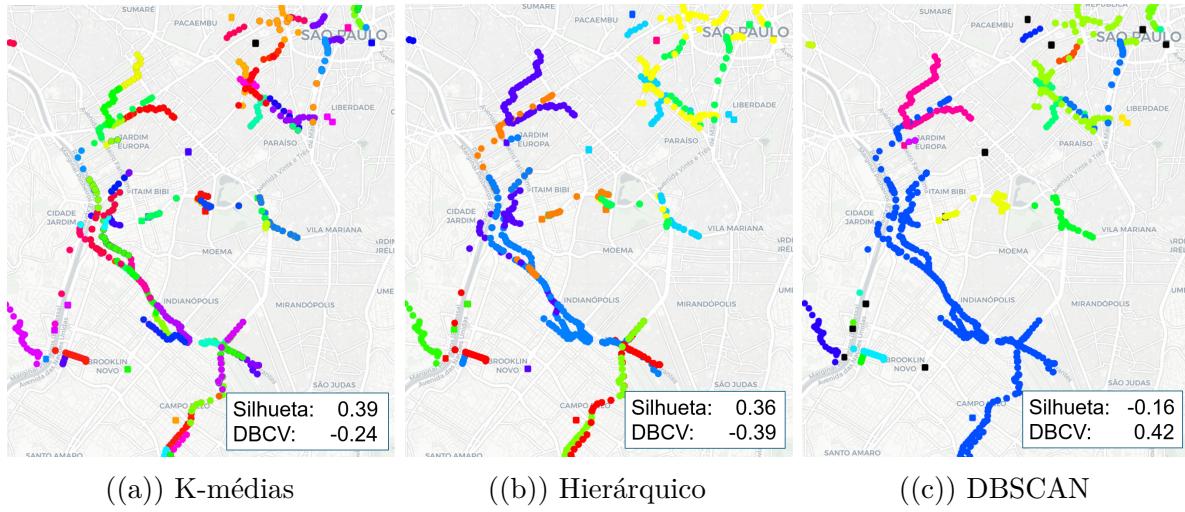


Figura 13 – Comparação dos algoritmos de clusterização para o dataset de arestas

Além da inspeção visual, foram comparadas as métricas de Silhueta e DBCV, detalhadas na seção 2.2.3. Observa-se que o coeficiente de Silhueta atribui valores mais altos para agrupamentos com formas esféricas, enquanto o DBCV favorece agrupamentos mais densos. A Figura 14 apresenta os valores de ambos os coeficientes para todas as sextas-feiras de 2019.

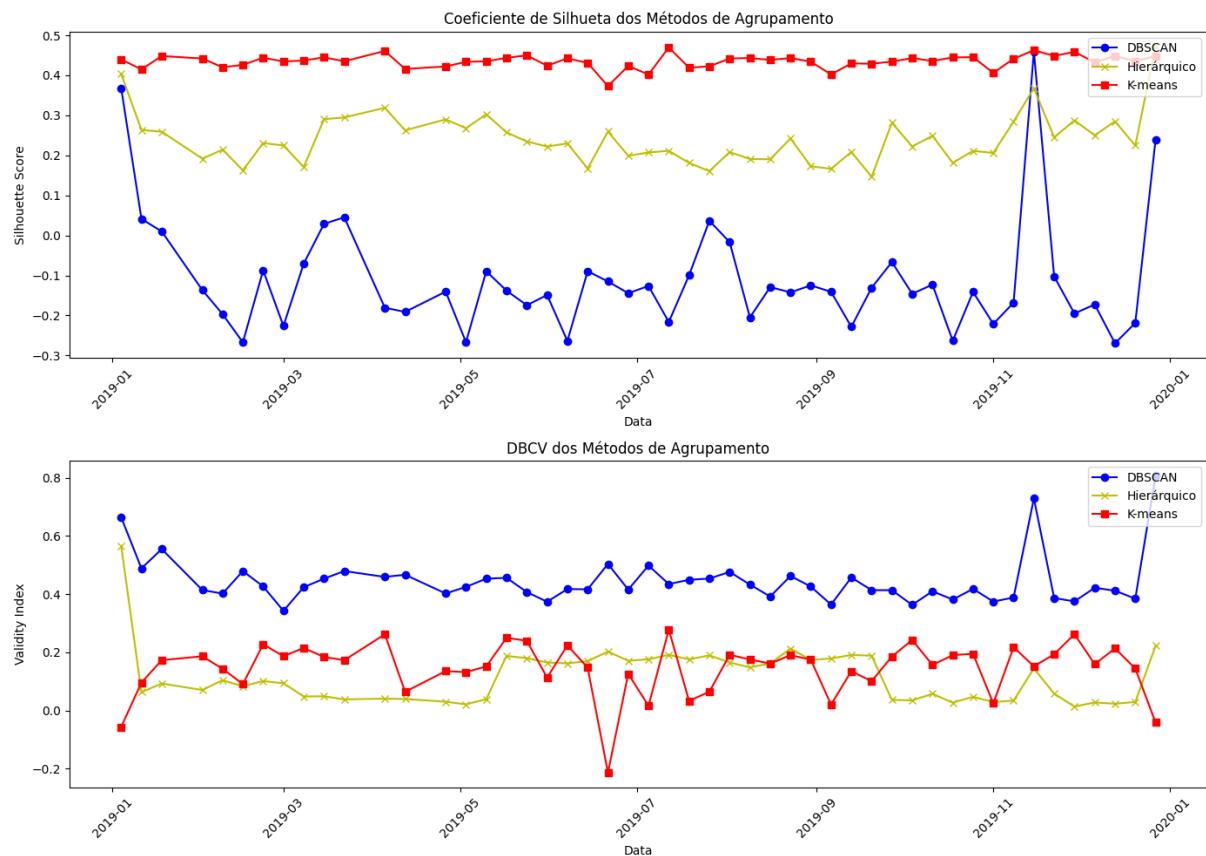


Figura 14 – Coeficiente de Silhueta e DBCV para os agrupamentos gerados pelos 3 algoritmos utilizados

Essa amostra foi escolhida para reduzir a complexidade do gráfico, tornar a geração dos resultados mais eficiente e evitar *outliers*, como os dias com poucos registros (feriados e finais de semana), nos quais os algoritmos se comportam de maneira semelhante. Ao comparar os índices, nota-se que, enquanto o coeficiente de Silhueta aponta o K-Médias como o melhor método de agrupamento, o DBCV indica o DBSCAN como o de melhor desempenho. Considerando o contexto específico do problema e as análises visuais, optou-se pelos resultados do algoritmo DBSCAN.

## 4.2 Características dos Eventos

A seguir, realizamos uma análise exploratória para identificar padrões nos dados. O dataset de eventos, resultante do agrupamento das 5 milhões de arestas, contém 83 mil registros. Primeiramente, comparamos a evolução mensal dos eventos (Figura 15).

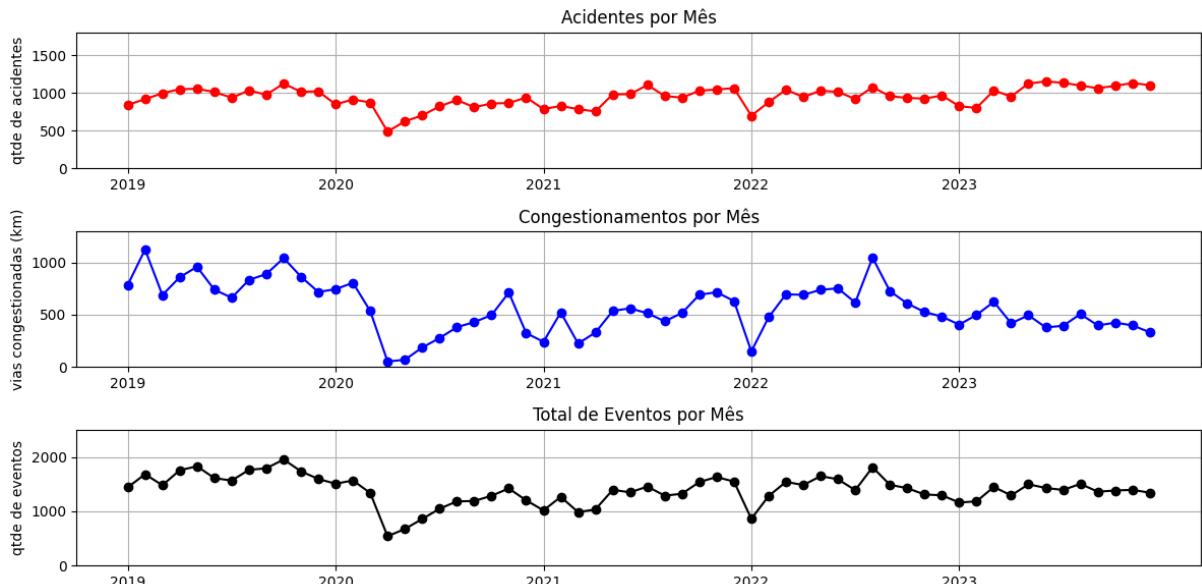


Figura 15 – Evolução mensal dos eventos

Na figura, pode-se observar o impacto das restrições de mobilidade impostas durante a pandemia de COVID-19, a partir de março de 2020. Nesse momento, o volume de congestionamentos reduziu quase por completo, enquanto a diminuição dos acidentes foi menos intensa. Notamos ainda que a quantidade de acidentes retomou os níveis de 2019 em poucos meses, enquanto a taxa de congestionamentos não alcançou os níveis pré-pandêmicos. É possível perceber, também, a regularidade da quantidade de acidentes, com cerca de 1000 acidentes a cada mês. Por outro lado, os congestionamentos possuem uma característica sazonal, em que o volume de um mês pode diferir significativamente do mês anterior. A sazonalidade também acontece nas horas do dia (Figura 16), onde os congestionamentos possuem dois períodos de pico (manhã e tarde), enquanto os acidentes têm menos variância ao longo do dia. Esses fatos podem indicar que os congestionamentos

não afetam significativamente os níveis de acidente, já que a quantidade de sinistros mantém-se regular mesmo com a variação do trânsito.

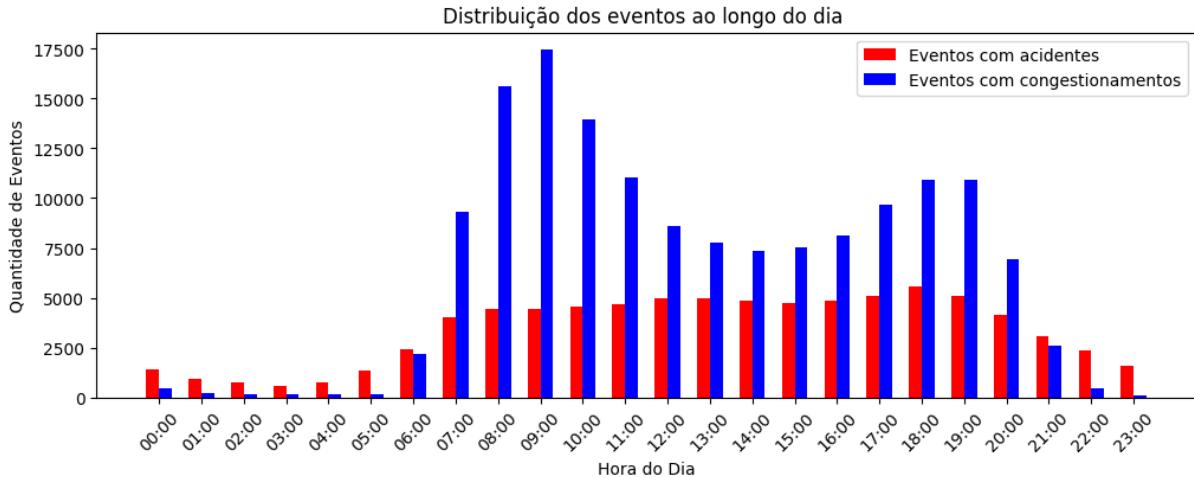


Figura 16 – Distribuição dos eventos ao longo do dia

Com relação à características dos congestionamentos a Figura 17 mostra a distribuição da duração e do tamanho dos trechos de lentidão. As vias congestionadas possuem, majoritariamente, menos de 2 quilômetros de extensão. A duração dos registros concentra-se em cerca de 2 horas, com uma segunda concentração de eventos com 12 horas de duração, que são ocorrências de congestionamentos que não se dissipam completamente entre os períodos de pico da manhã e da tarde.

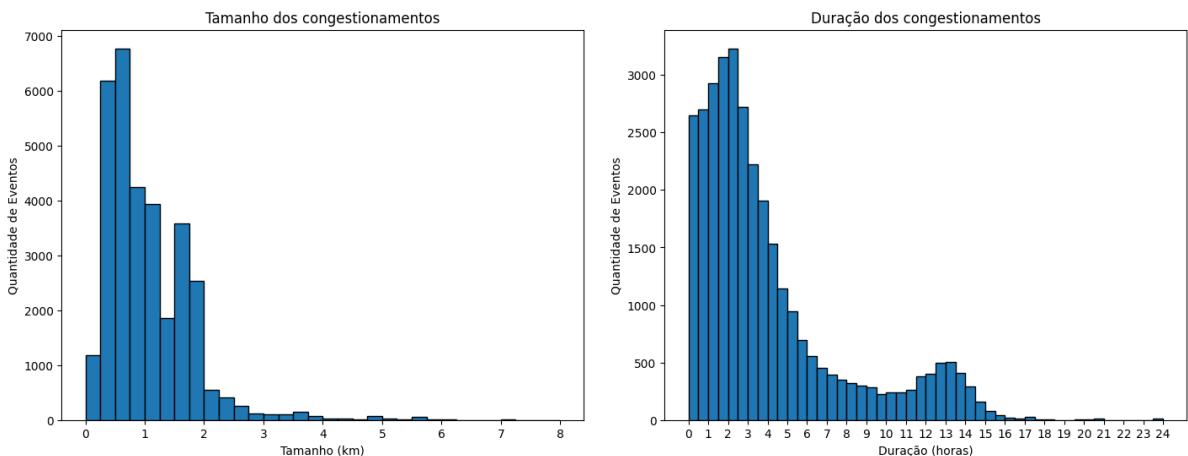


Figura 17 – Duração e tamanho dos congestionamentos

As características dos eventos envolvendo acidentes são ilustradas na Figura 18. Observa-se que acidentes graves são mais raros, assim como eventos que combinam congestionamentos e acidentes simultâneos, representando apenas 5% do total. A maioria dos acidentes ocorre sem a presença de congestionamentos, da mesma forma que a maioria dos congestionamentos não está associada a sinistros. Essa característica sugere uma correlação inversa entre a lentidão no trânsito e a ocorrência de acidentes, como se a

presença de congestionamentos diminuísse a probabilidade de colisões ou atropelamentos. Esse aspecto será analisado de forma mais detalhada com o uso dos modelos de regressão na próxima seção.

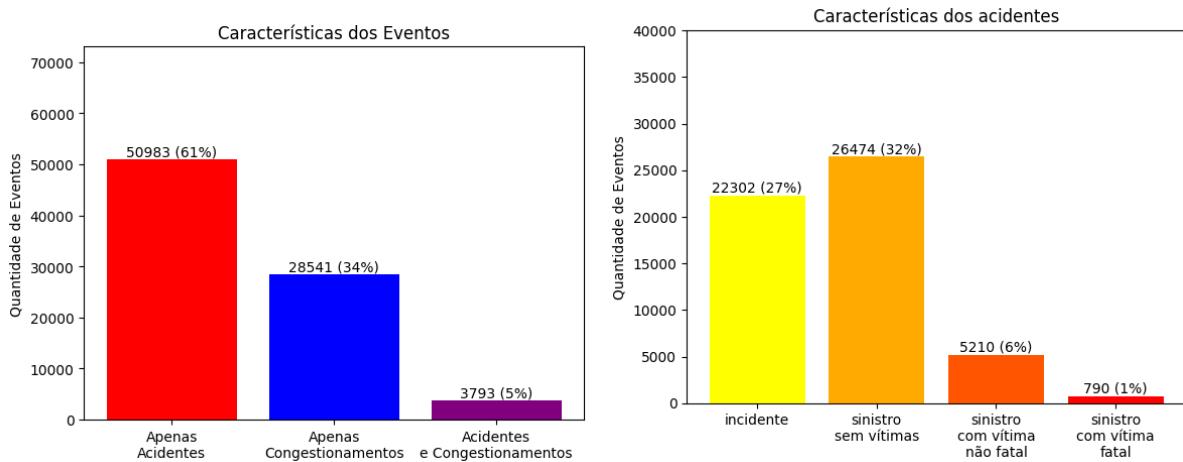


Figura 18 – Características dos eventos com acidentes

### 4.3 Modelos de Regressão

Conforme detalhado no capítulo Metodologia (3), aplicamos modelos de Regressão Polinomial utilizando a gravidade do acidente como variável dependente. Obtivemos os resultados indicados na Tabela 1. Os valores das métricas MSE e  $R^2$  foram calculados com o conjunto de testes (20%).

Grau	MSE	$R^2$	coeficiente mais significativo	valor do coeficiente
1	0.081	0.281	Extensão do Congestionamento	-0.203
2	0.064	0.405	Extensão do Congestionamento	-0.209
3	0.062	0.456	Extensão do Congestionamento	-0.315
4	0.049	0.476	Extensão do Congestionamento	-0.303

Tabela 1 – Comparação entre os modelos de regressão polinomial

Com os resultados, podemos observar que a regressão linear (grau 1) possuiu maior erro e menor coeficiente de determinação, ou seja, não foi capaz de explicar grande parte da variabilidade da variável dependente. À medida que o grau da regressão aumenta, o erro quadrático médio diminui e o coeficiente de determinação ( $R^2$ ) aumenta. Com grau 4, o modelo é capaz de determinar, com boa precisão, a variação da gravidade dos acidentes. Aproximadamente 47% da variância total dos dados é explicada pelos dados de congestionamentos e topologia das vias. O restante da variância é atribuído a fatores que não estão contemplados no modelo, como qualidade das vias e hábitos culturais.

Ao compararmos o papel de cada variável independente, observamos um comportamento comum em todas as implementações: o impacto do **tamanho do congestionamento**. Tanto para a regressão linear (Tabela ??), quanto para a regressão polinomial de

variável	coeficiente
intercepto	0.000000
duracao	-0.052882
dia_semana	0.020539
hora_float	0.011226
<b>tamanho</b>	<b>-0.203204</b>
<b>n_vertices</b>	<b>0.123364</b>
media_grau	-0.006877

Tabela 2 – Coeficientes da variáveis independentes no modelo de regressão linear

grau 2 (Figura 19) a extensão do congestionamento foi a variável com coeficiente mais significativo para determinar a gravidade do acidente. Em todos os casos, o valor foi negativo, indicando uma **correlação inversa**, ou seja, quanto menor o congestionamento, maior a gravidade do acidente.

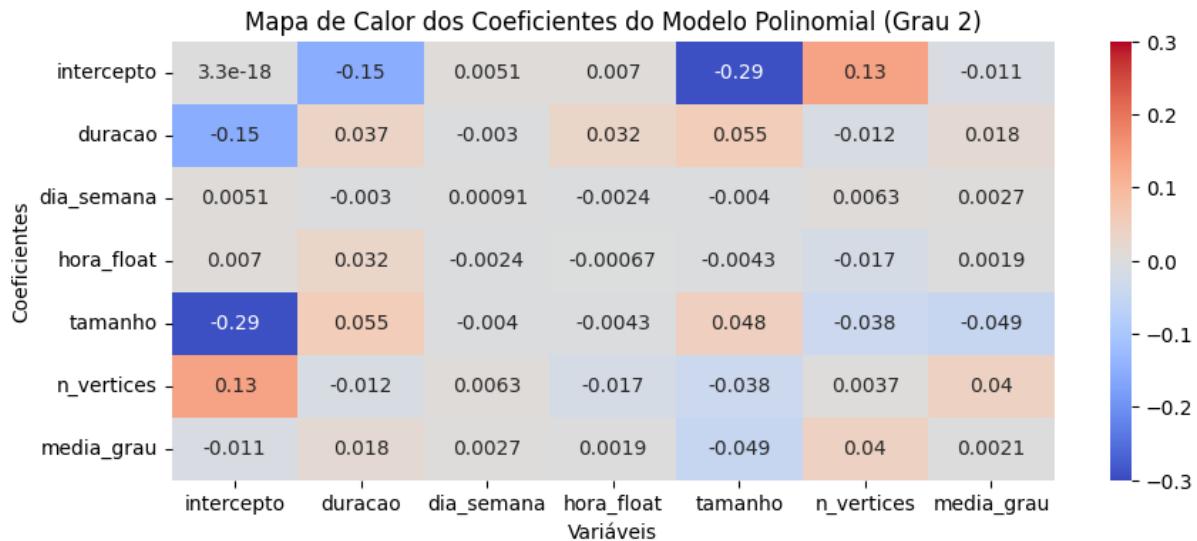


Figura 19 – Coeficientes das variáveis independentes no modelo de regressão polinomial de grau 2

Nas figuras 20 e 21, são exibidas projeções em 2D do comportamento da função modelada para cada variável independente. Nessas projeções, observa-se que a gravidade dos acidentes aumenta conforme o tamanho do congestionamento se aproxima de zero. Outro atributo que se destaca é a quantidade de vértices (número de interseções e curvas nas vias); quanto maior o número de vértices, maior a propensão à ocorrência de acidentes.

Em suma, os modelos de regressão aplicados neste estudo identificaram uma correlação inversa significativa entre congestionamentos e acidentes. No capítulo seguinte, discutiremos as causas e implicações desse resultado, além de sugerir ações e trabalhos futuros que possam enriquecer as conclusões apresentadas.

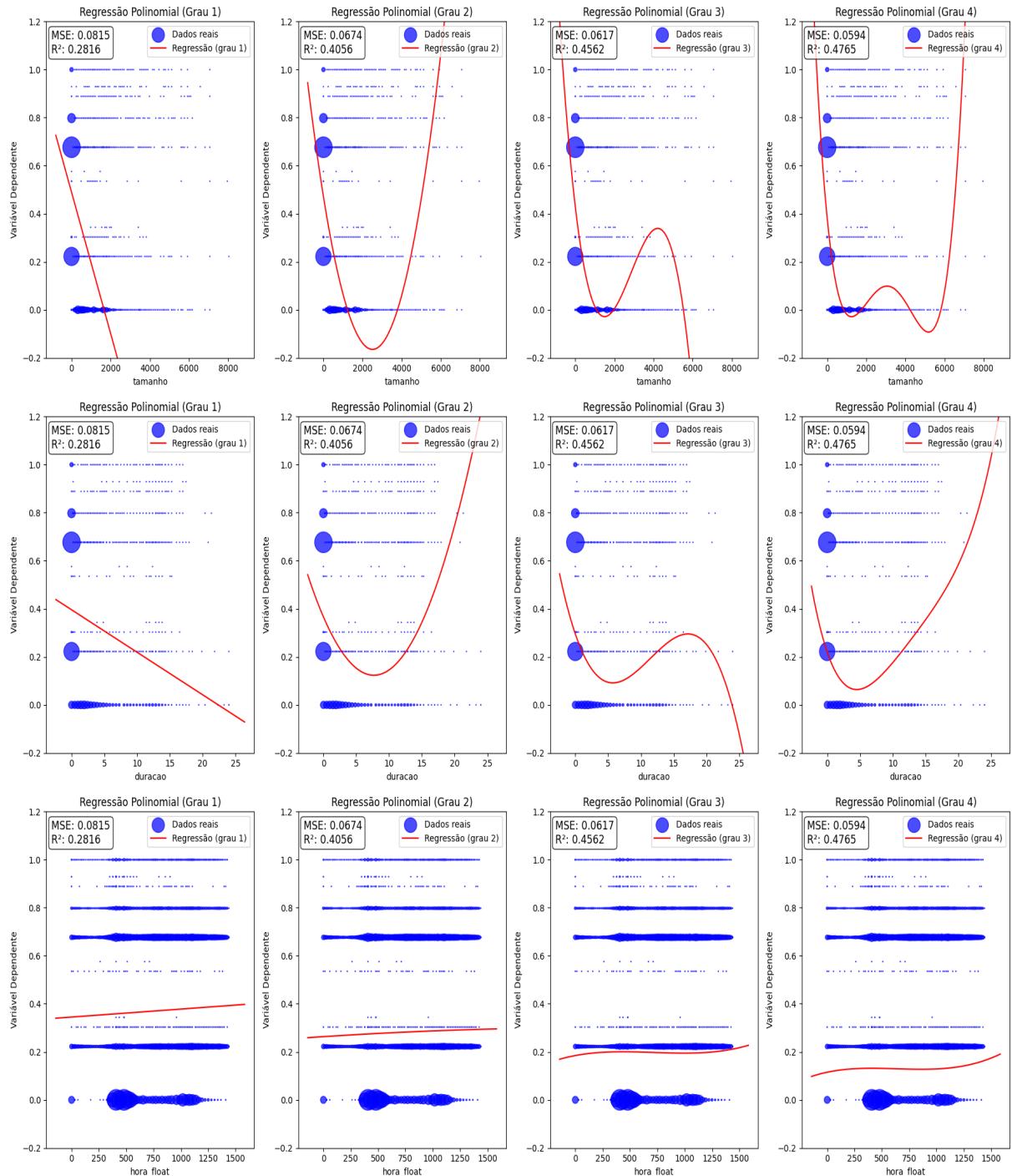


Figura 20 – Projeção da função de regressão modelada para cada variável independente (tamanho, duração, hora do dia)

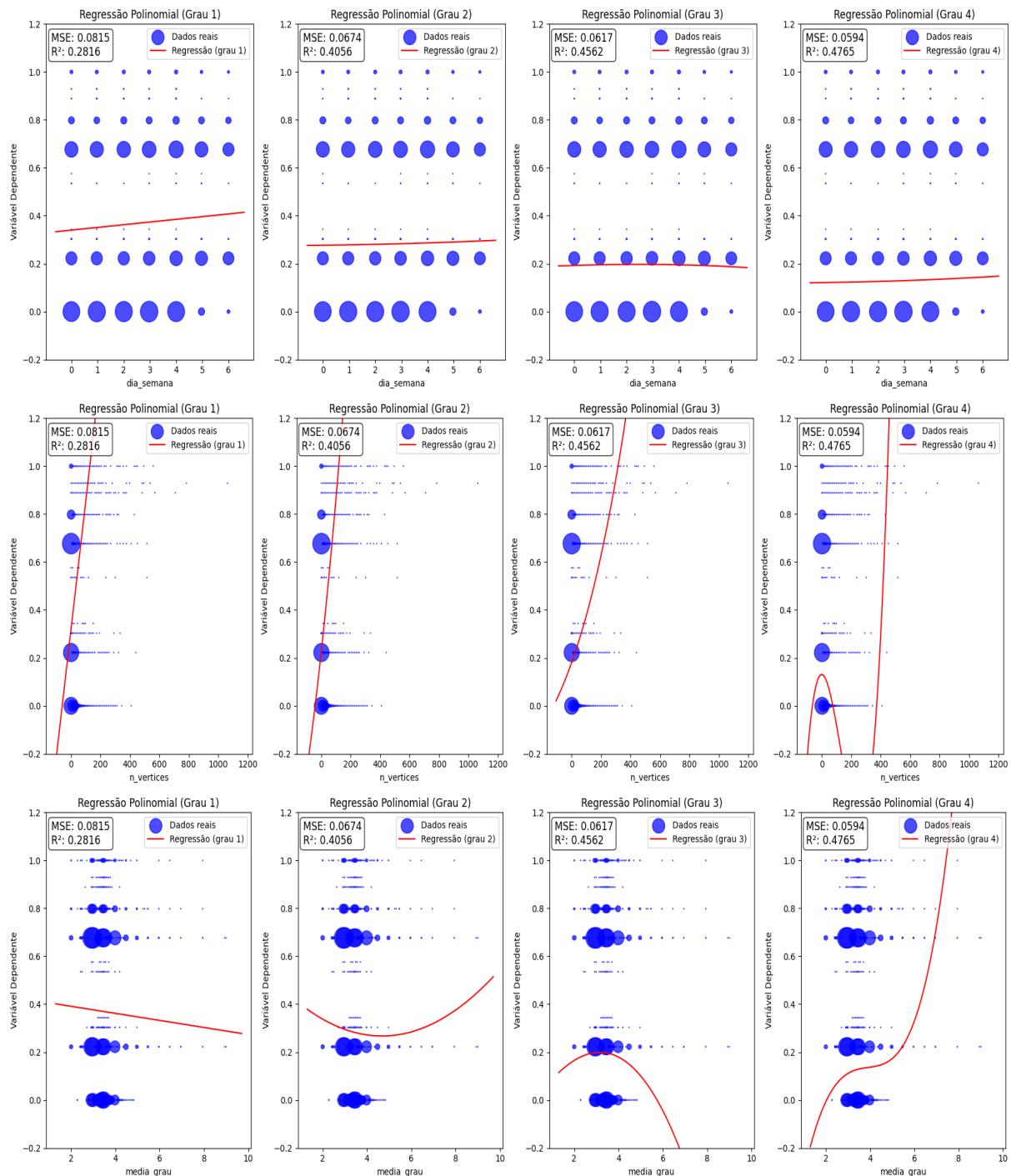


Figura 21 – Projeção da função de regressão modelada para cada variável independente (dia da semana, número de vértices e média do grau dos vértices)

## 5 CONCLUSÕES

Este trabalho teve como principal objetivo extrair conhecimento sobre a relação entre congestionamentos e acidentes de trânsito. Ao longo dos capítulos anteriores, foram discutidas diferentes abordagens de agrupamento de dados, os desafios envolvidos e as métricas de avaliação utilizadas para comparar o desempenho dos algoritmos. A partir da métrica DBCV, o algoritmo DBSCAN destacou-se como a melhor estratégia para agrupar os registros estudadas.

A metodologia aplicada para o agrupamento de registros dispostos em uma estrutura de grafo mostrou-se eficaz e escalável. Trabalhos futuros podem explorar essa estrutura de dados para incluir novas informações que contribuem para o modelo. As arestas do grafo podem incorporar características das vias, como largura, aclives e declives, presença de semáforos, faixas de pedestres, entre outros. Além disso, a estrutura de dados permite a associação com informações de mobilidade urbana, como linhas de ônibus e metrô, bem como dados socioeconômicos, como densidade populacional e postos de trabalho.

Os autores esperam que a inclusão de novos dados possibilite a criação de modelos capazes de indicar com precisão a relevância de outras causas de acidentes de trânsito. Dessa forma, as autoridades de trânsito poderão atuar de maneira mais eficaz na mitigação dos prejuízos causados por esses acidentes, adotando uma abordagem orientada por dados e evidências científicas (*data-driven*).

No que diz respeito aos congestionamentos, os modelos de regressão indicaram, com boa significância estatística, que **os congestionamentos reduzem a intensidade dos acidentes de trânsito**. No entanto, essa conclusão parece ocultar a importância de uma variável implícita: a **velocidade dos veículos**. Quanto maior e mais prolongado é o congestionamento, mais lentamente os veículos trafegam. Este trabalho propõe, portanto, que a velocidade nas vias seja regulada de forma a aumentar a segurança no trânsito.

A redução da velocidade não se limita apenas à imposição de limites máximos e à fiscalização correspondente. Outros fatores, como a largura das faixas, a sincronia dos semáforos, a sinalização de curvas e obstáculos, além da qualidade dos materiais utilizados na pavimentação, também desempenham um papel crucial na organização do fluxo de veículos, contribuindo para a prevenção de acidentes (Aldegeishem *et al.*, 2018).

Apesar da correlação inversa entre os congestionamentos e acidentes, os coeficientes dessa correlação apresentam baixa magnitude. Em outras palavras, embora exista correlação, ela não é forte, já que acidentes ocorrem em diferentes níveis de congestionamento. Reforçando essa observação, o coeficiente de determinação dos modelos ficou em torno de 40%, o que sugere que a solução para o problema dos acidentes de trânsito deve considerar

mais dados e incluir análises mais específicas, como o estudo de regiões e vias particulares

Por fim, este estudo implementou uma metodologia robusta e escalável para extrair conhecimento de dados georreferenciados de acidentes e congestionamentos. Identificamos uma correlação inversa entre acidentes e a lentidão dos veículos, oferecendo sugestões para a implementação de ações que mitiguem o risco de incidentes graves. Com isso, contribuímos para o avanço de um tema ainda sem consenso na literatura especializada. Trabalhos futuros podem incorporar novas fontes de dados, especialmente em um cenário de *Big Data*, onde a democratização e o crescente volume de dados permitirão explorar variáveis adicionais e aumentar a precisão das conclusões. Orientar a segurança no trânsito por meio de tecnologias como a ciência de dados e a inteligência artificial oferece oportunidades reais de evoluir a qualidade da mobilidade urbana em países em desenvolvimento, salvando vidas que se perdem diariamente em acidentes.

## REFERÊNCIAS

- ABNT. Abnt nbr 10697. pesquisa de sinistros de trânsito — terminologia. **Associação Brasileira de Normas Técnicas**, Rio de Janeiro, RJ, 2020. Disponível em: <https://www.abramet.com.br/repo/public/commons/ABNT%20NBR10697%202020%20Acidentes%20de%20Transito%20Terminologia.pdf>. Acesso em: 07 abr. 2024.
- AGGARWAL, C. C. **Data Mining: The Textbook**. Springer International Publishing, 2015. ISBN 9783319141428. Disponível em: <http://dx.doi.org/10.1007/978-3-319-14142-8>.
- ALDEGHEISHEM, A. *et al.* Smart road traffic accidents reduction strategy based on intelligent transportation systems (tars). **Sensors**, v. 18, n. 7, 2018. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/18/7/1983>.
- ALMUKHALFI, H.; NOOR, A.; NOOR, T. H. Traffic management approaches using machine learning and deep learning techniques: A survey. **Engineering Applications of Artificial Intelligence**, Elsevier BV, v. 133, p. 108147, jul. 2024. ISSN 0952-1976. Disponível em: <http://dx.doi.org/10.1016/j.engappai.2024.108147>.
- BACCHIERI, G.; BARROS, A. J. D. Acidentes de trânsito no brasil de 1998 a 2010: muitas mudanças e poucos resultados. **Revista de Saúde Pública**, FapUNIFESP (SciELO), v. 45, n. 5, p. 949–963, out. 2011. ISSN 0034-8910. Disponível em: <http://dx.doi.org/10.1590/S0034-89102011005000069>.
- BARIONI, M. C. N. Operações de consulta por similaridade em grandes bases de dados complexos. **São Carlos, SP: Universidade de São Paulo**, 2006.
- BATISTA, D. M. *et al.* Intercity: Addressing future internet research challenges for smart cities. In: **2016 7th International Conference on the Network of the Future (NOF)**. [S.l.: s.n.], 2016. p. 1–6.
- BRASIL. Lei nº 9.503, de 23 de setembro de 1997. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 1997. Disponível em: [https://www.planalto.gov.br/ccivil\\_03/\\_ato2011-2014/2012/lei/l12587.htm](https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12587.htm). Acesso em: 07 abr. 2024.
- BRASIL. Lei nº 12.587, de 3 de janeiro de 2012. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2012. Disponível em: [https://www.planalto.gov.br/ccivil\\_03/\\_ato2011-2014/2012/lei/l12587.htm](https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12587.htm). Acesso em: 07 abr. 2024.
- BUREAU OF INFRASTRUCTURE, TRANSPORT AND REGIONAL ECONOMICS. **Impact of road trauma and measures to improve outcomes**. Canberra, 2014.
- CALATAYUD, A. *et al.* **Congestión urbana en América Latina y el Caribe: Características, costos y mitigación**. Inter-American Development Bank, 2021. Disponível em: <http://dx.doi.org/10.18235/0003149>.
- CET-SP. Acidentes de trânsito relatório anual 2021. São Paulo, SP, 2022. Disponível em: <https://www.cetsp.com.br/media/1347066/Relatorioanual2021.pdf>. Acesso em: 28 set. 2024.

CET-SP. **Base de dados sobre lentidão por trechos - CET**. 2024. Disponível em: <http://dados.prefeitura.sp.gov.br/it/dataset/base-de-dados-sobre-lentidao-por-trechos-cet>.

CHAND, A.; JAYESH, S.; BHASI, A. Road traffic accidents: An overview of data sources, analysis techniques and contributing factors. **Materials Today: Proceedings**, v. 47, p. 5135–5141, 2021. ISSN 2214-7853. International Conference on Sustainable materials, Manufacturing and Renewable Technologies 2021. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2214785321040153>.

CHAUDHARY, R.; BHADAURIA, M.; GARG, A. Predictive analysis techniques for road accident injuries: A survey. In: **2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)**. [S.l.: s.n.], 2017. p. 91–96.

CHAWLA, A. **The Limitation Of Silhouette Score Which Is Often Ignored By Many**. 2023. Disponível em: <https://blog.dailydoseofds.com/p/the-limitation-of-silhouette-score>.

ESTER, M. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: **kdd**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FEOFILOFF, P.; YOSHIKO, W.; KOHAYAKAWA, Y. **Uma Introdução Sucinta à Teoria dos Grafos**. [S.l.: s.n.], 2011.

GONZÁLEZ, S. S.; BEDOYA-MAYA, F.; CALATAYUD, A. Understanding the effect of traffic congestion on accidents using big data. **Sustainability**, MDPI AG, v. 13, n. 13, p. 7500, jul. 2021.

GOODWIN, P. The economic costs of road traffic congestion. UCL (University College London), The Rail Freight Group, 2004.

GOYAL, A. **Learning a Multiview Weighted Majority Vote Classifier: Using PAC-Bayesian Theory and Boosting**. 10 2018. Tese (Doutorado), 10 2018.

HASSAN, B. A. *et al.* From a-to-z review of clustering validation indices. **Neurocomputing**, v. 601, p. 128198, 2024. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S092523122400969X>.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. Springer US, 2021. ISSN 2197-4136. ISBN 9781071614181. Disponível em: <http://dx.doi.org/10.1007/978-1-0716-1418-1>.

JARMAN, A. M. Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. **Georgia Southern University**, v. 29, 2020.

KAUFMAN, L.; ROUSSEEUW, P. **Finding Groups in Data: An Introduction To Cluster Analysis**. [S.l.: s.n.], 1990. ISBN 0-471-87876-6.

- LARA, W. Número de mortes em acidentes no trânsito de SP em 2023 é o maior dos últimos oito anos. 2024. Disponível em: <https://g1.globo.com/sp/sao-paulo/noticia/2024/01/16/numero-de-mortes-em-acidentes-no-transito-de-sp-em-2023-e-o-maior-dos-ultimos-oito-anos.ghtml>.
- LI, X.; GUI, J.; LIU, J. Data-driven traffic congestion patterns analysis: a case of beijing. **Journal of Ambient Intelligence and Humanized Computing**, Springer Science and Business Media LLC, v. 14, n. 7, p. 9035–9048, set. 2022. ISSN 1868-5145. Disponível em: <http://dx.doi.org/10.1007/s12652-022-04409-4>.
- LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, n. v. 4, n. 4, p. 18–36, 2009.
- MALATESTA, M. E. B. **A bicicleta nas viagens cotidianas do Município de São Paulo**. 2014. Tese (Doutorado) — Faculdade de Arquitetura e Urbanismo, Universidade de São Paulo, 2014.
- MARTIN, J.-L. *et al.* Cannabis, alcohol and fatal road accidents. **PLoS One**, Public Library of Science (PLoS), v. 12, n. 11, p. e0187320, nov. 2017.
- MOHAMMED, A. A. *et al.* A review of the traffic accidents and related practices worldwide. **The Open Transportation Journal**, Bentham Science Publishers Ltd., v. 13, n. 1, p. 65–83, jun. 2019. ISSN 1874-4478. Disponível em: <http://dx.doi.org/10.2174/1874447801913010065>.
- MONTGOMERY ELIZABETH A. PECK, G. G. V. D. C. **Introduction to Linear Regression Analysis, Fifth Edition**. Wiley, 2013. v. 81. 318–319 p. ISSN 1751-5823. Disponível em: [http://dx.doi.org/10.1111/insr.12020\\_10](http://dx.doi.org/10.1111/insr.12020_10).
- MOULAVI, D. *et al.* Density-based clustering validation. In: SIAM. **Proceedings of the 2014 SIAM international conference on data mining**. [S.l.: s.n.], 2014. p. 839–847.
- NIELSEN, F. **Introduction to HPC with MPI for Data Science**. Springer International Publishing, 2016. ISSN 2197-1781. ISBN 9783319219035. Disponível em: <http://dx.doi.org/10.1007/978-3-319-21903-5>.
- PYTHON. **difflib — Helpers for computing deltas**. 2023. Disponível em: <https://docs.python.org/3/library/difflib.html>.
- QUDDUS, M. A.; WANG, C.; ISON, S. G. Road traffic congestion and crash severity: Econometric analysis using ordered response models. **Journal of Transportation Engineering**, American Society of Civil Engineers (ASCE), v. 136, n. 5, p. 424–435, maio 2010. ISSN 1943-5436. Disponível em: [http://dx.doi.org/10.1061/\(ASCE\)TE.1943-5436.0000044](http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000044).
- RETALLACK, A. E.; OSTENDORF, B. Current understanding of the effects of congestion on traffic accidents. **Int. J. Environ. Res. Public Health**, MDPI AG, v. 16, n. 18, p. 3400, set. 2019.
- RHYS, H. **Machine Learning with R, the tidyverse, and mlr**. New York, NY: Manning Publications, 2020.

- ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>.
- RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: a modern approach**. 3. ed. [S.l.: s.n.]: Pearson, 2009.
- SCIPY. **KDTree**. 2024. Disponível em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.html>.
- SCYKIT. **StandardScaler**. 2024. Disponível em: <https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.StandardScaler.html>.
- SHAPELY. **shapely.distance**. 2024. Disponível em: <https://shapely.readthedocs.io/en/stable/reference/shapely.distance.html>.
- SHELLER, M.; URRY, J. The city and the car. **International Journal of Urban and Regional Research**, v. 24, n. 4, p. 737–757, 2000. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-2427.00276>.
- UFRN. **Regressão Polinomial**. 2018. Disponível em: <https://cn.ect.ufrn.br/index.php?r=conteudo%2Fmmq-rpolin>.
- VASCONCELLOS, E. A. Urban development and traffic accidents in brazil. **Accident Analysis & Prevention**, Elsevier BV, v. 31, n. 4, p. 319–328, jul. 1999. ISSN 0001-4575. Disponível em: [http://dx.doi.org/10.1016/S0001-4575\(98\)00065-7](http://dx.doi.org/10.1016/S0001-4575(98)00065-7).
- WORLD HEALTH ORGANIZATION. **Road traffic injuries**. [S.l.], 2023. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.
- ZOHRA, E. F. *et al.* Accident severity prediction using machine learning: A case study on the us accidents dataset. In: **2023 17th International Conference on Signal-Image Technology; Internet-Based Systems (SITIS)**. IEEE, 2023. Disponível em: <http://dx.doi.org/10.1109/SITIS61268.2023.00044>.