INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO
PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA APLICADA E SUSTENTABILIDADE - MESTRADO PROFISSIONAL
CAMPUS RIO VERDE

ANÁLISE EXPLORATÓRIA E ESPACIAL DE SINISTROS DE TRÂNSITO NA ÁREA URBANA DE RIO VERDE (GO) COM UTILIZAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Autor: Michel Gondim Oliveira
Orientador: Philippe Barbosa Silva
Coorientador: Geraldo Andrade de Oliveira

RIO VERDE - GO
dezembro - 2025

MICHEL GONDIM OLIVEIRA

ANÁLISE EXPLORATÓRIA E ESPACIAL DE SINISTROS DE TRÂNSITO NA ÁREA URBANA DE RIO VERDE (GO) COM UTILIZAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Dissertação apresentada à banca examinadora como parte das exigências para obtenção do título de Mestre em Engenharia Aplicada e Sustentabilidade, do Programa de Pós-Graduação em Engenharia Aplicada e Sustentabilidade do Instituto Federal de Educação, Ciência e Tecnologia Goiano – Campus Rio Verde – Linha de Pesquisa I – Tecnologia e gestão em construção civil e infraestrutura

Orientador: Prof. Dr. Philippe Barbosa Silva
Coorientador: Prof. Dr. Geraldo Andrade de Oliveira
.

RIO VERDE, GO
dezembro – 2025

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema Integrado de Bibliotecas do IF Goiano - SIBi

O48a

Oliveira, Michel Gondim

ANÁLISE EXPLORATÓRIA E ESPACIAL DE SINISTROS DE TRÂNSITO NA ÁREA URBANA DE RIO VERDE (GO) COM UTILIZAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA / Michel Gondim Oliveira. Rio Verde 2025.

120f. il.

Orientador: Prof. Dr. Philippe Barbosa Silva.
Coorientador: Prof. Dr. Geraldo Andrade de Oliveira.
Dissertação (Mestre) - Instituto Federal Goiano, curso de 0233144 - Mestrado Profissional em Engenharia Aplicada e Sustentabilidade (Campus Rio Verde).

1. Sinistros de trânsito. 2. Análise espacial. 3. Aprendizado de máquina. 4. Rio Verde. 5. Segurança viária.. I. Título.

INSTITUTO FEDERAL
Goiano

Repositório Institucional do IF Goiano - RIIF Goiano
Sistema Integrado de Bibliotecas

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☑ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☐ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor: MICHEL GONDIM OLIVEIRA

Matrícula: 2023102331440004

Título do trabalho:

ANÁLISE EXPLORATÓRIA E ESPACIAL DE SINISTROS DE TRÂNSITO NA ÁREA URBANA DE RIO VERDE (GO) COM UTILIZAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☐ Não ☑ Sim, justifique:

O conteúdo resultará na submissão de artigos científicos a revistas nacionais e internacionais.

Informe a data que poderá ser disponibilizado no RIIF Goiano: 30 / 06 / 2026

O documento está sujeito a registro de patente? ☑ Sim ☐ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

• Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

• Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

• Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

gov.br Documento assinado digitalmente
MICHEL GONDIM OLIVEIRA
Data: 16/12/2025 14:15:47-0300
Verifique em https://validar.iti.gov.br

RIO VERDE, GO — Local

16 / 12 / 2025 — Data

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

gov.br Documento assinado digitalmente
PHILIPPE BARBOSA SILVA
Data: 16/12/2025 14:21:32-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) orientador(a)

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Documentos 54/2025 - SREPG/CMPR/CPG-RV/DPGPI-RV/CMPRV/IFGOIANO

ANÁLISE EXPLORATÓRIA E ESPACIAL DE SINISTROS DE TRÂNSITO NA ÁREA URBANA DE RIO VERDE (GO) COM UTILIZAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Autor: Michel Gondim Oliveira
Orientador: Prof. Dr. Philippe Barbosa Silva

TITULAÇÃO: Mestre em Engenharia Aplicada e Sustentabilidade - Área de Concentração Engenharia Aplicada e Sustentabilidade

APROVADO em 30 de junho de 2025.

Profª. Dra. Michelle Andrade
Avaliadora Externa - Universidade de Brasília

Profª. Dra. Sara Maria Pinho Ferreira
Avaliadora Externa - Universidade do Porto (Portugal)

Prof. Dr. Philippe Barbosa Silva
Presidente da banca - IFGOIANO / Rio Verde

Documento assinado eletronicamente por:

- **Philippe Barbosa Silva**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 08/09/2025 15:48:05.
- **Sara Maria Pinho Ferreira**, **Sara Maria Pinho Ferreira - Professor Avaliador de Banca - Instituto Federal Goiano - Campus Rio Verde (10651417000500)**, em 08/09/2025 16:45:26.
- **Michelle Andrade**, **Michelle Andrade - Professor Avaliador de Banca - Universidade de Brasília (00038174000143)** , em 09/09/2025 18:27:19.

Este documento foi emitido pelo SUAP em 27/05/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 711052
Código de Autenticação: 54fc298b7e

INSTITUTO FEDERAL GOIANO
Campus Rio Verde
Rodovia Sul Goiana, Km 01, Zona Rural, 01, Zona Rural, RIO VERDE / GO, CEP 75901-970
(64) 3624-1000

AGRADECIMENTOS

Com imensa gratidão, expresso meus sinceros agradecimentos a todos aqueles que, de alguma forma, contribuíram para a conclusão desta jornada acadêmica e pessoal.

Aos meus pais e irmãos, pelo apoio constante, pela paciência nas ausências inevitáveis e por acreditarem, desde o início, na importância dessa conquista.

Ao meu orientador, Professor Doutor Philippe Barbosa Silva, pelo apoio incondicional, pela compreensão nos momentos de dificuldade, pela orientação criteriosa e pela confiança que sempre depositou no meu trabalho. Sua escuta atenta e sua generosidade intelectual foram fundamentais para que este percurso fosse possível.

Ao meu coorientador, Professor Doutor Geraldo Andrade de Oliveira, por ter me incentivado a buscar esse almejado título, por ter apresentado o Mestrado do IF Goiano como uma possibilidade concreta, e por ter oferecido apoio desde os primeiros passos desta trajetória.

Aos professores Leonardo Garcia, Jesmmer e Josemar, pelo suporte ao longo dos semestres e pelas reuniões aos sábados pela manhã, momentos em que nosso grupo compartilhava avanços, dúvidas e aprendizados.

Ao Professor Leandro Souza, pelo suporte incondicional nesta reta final, cuja ajuda foi decisiva para a finalização desta dissertação.

Ao Programa de Pós-Graduação em Engenharia Aplicada e Sustentabilidade, agradeço a estrutura e pelos recursos disponibilizados, fundamentais para a realização desta pesquisa, e, em especial, ao coordenador do programa, Professor Doutor Édio Damásio da Silva Junior. Estendo meus agradecimentos a todos os docentes das disciplinas cursadas e à dedicada equipe de secretaria e apoio do IF Goiano – Campus Rio Verde.

A todos os mencionados e àqueles que, direta ou indiretamente, contribuíram para este momento, meu mais profundo obrigado. Cada gesto de apoio foi fundamental para o sucesso desta conquista. Estou eternamente grato por fazerem parte desta trajetória significativa em minha vida.

BIOGRAFIA DO AUTOR

Michel Gondim Oliveira é natural de Fortaleza (CE). Doutor Honoris Causa em Educação Profissional e Tecnológica, é mestrando em Engenharia Aplicada e Sustentabilidade pelo Instituto Federal Goiano – Campus Rio Verde e graduado em Ciências Econômicas pela Universidade Federal do Tocantins. Possui experiência em gestão de projetos públicos e como consultor na iniciativa pública e privada, com atuação como coordenador, assessor técnico e gerente de projetos.

Atuou diretamente na implementação de iniciativas educacionais voltadas à capacitação profissional e à inserção no mercado de trabalho, com atendimento direto a mais de 9.500 alunos em diversos estados do país. Possui trajetória em projetos de pesquisa, diagnósticos territoriais, capacitação técnica, desenvolvimento socioprodutivo e articulação institucional. Tem experiência na elaboração de projetos internacionais e atuou como consultor pela *Deutsche Gesellschaft für Internationale Zusammenarbeit* (GIZ) GmbH, no âmbito da cooperação técnica Brasil-Alemanha e pelo Instituto Interamericano de Cooperação para a Agricultura (IICA).

Representou o estado do Tocantins em eventos temáticos da indústria promovidos pela Confederação Nacional da Indústria (CNI) e colaborou com projetos voltados à gestão fundiária e ambiental no Programa Nacional de Reforma Agrária (PNRA).

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

ÍNDICE DE QUADROS

ÍNDICE DE SÍMBOLOS, SIGLAS E ABREVIAÇÕES

ABNT	Associação Brasileira de Normas Técnicas
AMT	Agência Municipal de Mobilidade e Trânsito
CNM	Confederação Nacional dos Municípios
CSV	Comma-Separated Values
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNIT	Departamento Nacional de Infraestrutura de Transportes
EPSG	European Petroleum Survey Group (código de referência cartográfica)
ESG	Environmental, Social and Governance
GIS	Geographic Information System
GO	Goiás
HTML	HyperText Markup Language
IA	Inteligência Artificial
IF GOIANO	Instituto Federal de Educação, Ciência e Tecnologia Goiano
KDE	Kernel Density Estimation (Estimativa de Densidade por *Kernel*)
LGPD	Lei Geral de Proteção de Dados Pessoais
LISA	Local Indicators of Spatial Association
ML	*Machine Learning* (Aprendizado de Máquina)
ODS	Objetivos de Desenvolvimento Sustentável
OMS	Organização Mundial da Saúde
ONU	Organização das Nações Unidas
PCA	Principal Component Analysis
PNATRANS	Plano Nacional de Redução de Mortes e Lesões no Trânsito
PPGEAS	Programa de Pós-Graduação em Engenharia Aplicada e Sustentabilidade
QGIS	Quantum GIS
R^2	Coeficiente de Determinação
RENAEST	Relatório Estatístico de Sinistros de Trânsito
SENATRAN	Secretaria Nacional de Trânsito
SUS	Sistema Único de Saúde
t-SNE - t-	Distributed Stochastic Neighbor Embedding
WGS	World Geodetic System

RESUMO

A dissertação investiga sinistros de trânsito ocorridos em Rio Verde (GO) entre 2021 e 2024, a partir de 7.926 boletins eletrônicos convertidos em bases adequadas para análise temporal, espacial e preditiva. A caracterização estatística indica estabilidade anual, pico vespertino em dias úteis, predominância de colisões bilaterais com automóveis e maior envolvimento de homens de 18 a 35 anos. Métodos de densidade *kernel*, *Moran* I/LISA e DBSCAN localizam focos persistentes na Avenida Presidente Vargas e nos eixos BR-452/GO-174, com autocorrelação espacial positiva que sugere trechos prioritários para intervenção. Modelos *Random Forest, LightGBM* e regressão logística, avaliados por validação cruzada estratificada e controle de vazamento temporal, alcançam acurácia entre 0,68 e 0,97 e mostram que variáveis geográficas, sazonais e de infraestrutura superam marcadores exclusivamente temporais na previsão de gravidade e características associadas ao sinistro. Conclui-se que a integração de estatística espacial e aprendizado de máquina tem potencial para orientar ações de engenharia, fiscalização e educação voltadas à redução de risco em corredores críticos.

Palavras-chave: sinistros de trânsito; análise espacial; aprendizado de máquina; Rio Verde; segurança viária.

ABSTRACT

The dissertation investigates the traffic crashes that occurred in Rio Verde (GO), Brazil, between 2021 and 2024, based on 7,926 electronic reports converted into datasets suitable for temporal, spatial, and predictive analysis. The statistical characterization indicates annual stability, afternoon peaks on weekdays, a predominance of bilateral collisions involving automobiles, and a higher incidence among men aged 18 to 35. Kernel density estimation, Moran's I/LISA, and DBSCAN methods identify persistent hotspots along Presidente Vargas Avenue and the BR-452/GO-174 corridors, with positive spatial autocorrelation suggesting priority areas for its intervention. Random Forest, LightGBM, and logistic regression models, evaluated through stratified cross-validation and temporal leakage control, achieved accuracies ranging from 0.68 to 0.97. Results show that geographic, seasonal, and infrastructure-related variables outperform purely temporal markers in predicting crash severity and associated characteristics. The study concludes

that integrating spatial statistics and machine learning supports engineering, enforcement, and educational strategies aimed at reducing risk in critical corridors.

Keywords: traffic crashes; spatial analysis; machine learning; Rio Verde; road safety.

1. INTRODUÇÃO

1.1. Justificativa

A segurança viária configura-se como um desafio global: estima-se que 1,19 milhão de pessoas morram anualmente em sinistros de trânsito, com maior concentração em países de renda média e baixa. Esse quadro motivou a Organização das Nações Unidas a instituir a Década de Ação pela Segurança no Trânsito 2021-2030, cuja meta principal é reduzir em 50% as mortes e lesões no período considerado (ONU, 2020).

No Brasil, os sinistros permanecem entre as principais causas externas de mortalidade; em 2019 foram registradas 32.654 mortes, produzindo impactos significativos sobre os sistemas de saúde e produtividade econômica (OMS, 2019). Em escala estadual, dados do Departamento de Informática do SUS indicam crescimento de 6,08% nas mortes por sinistros em Goiás, entre 2020 e 2021, passando de 1.578 para 1.674 óbitos (Brasil, 2023). Esse incremento reafirma a necessidade de políticas focalizadas que considerem especificidades regionais e municipais. Para o período 2021-2024, o Relatório Estatístico de Sinistros de Trânsito (RENAEST/SENATRAN) informa 15.824 registros em Rio Verde, envolvendo 22.523 veículos, resultando em 102 óbitos, correspondendo a 43,02 mortes por 100.000 habitantes (SENATRAN, 2024). Esses indicadores refletem a combinação de tráfego agroindustrial intenso, expansão urbana acelerada e hierarquia viária centralizada em poucos corredores, ampliando a exposição ao risco.

A literatura aponta que jovens do sexo masculino, usuários de motocicletas e condutores em fins de semana, à noite, formam o perfil mais frequente de vítimas (OMS, 2021). Ao mesmo tempo, fatores estruturais, como condição do pavimento, sinalização e geometria das vias, interagem de modo complexo com aspectos comportamentais, exigindo abordagem multifatorial para compreensão e mitigação do problema. Nesse cenário, a presente dissertação parte de diagnóstico de âmbito global e nacional, justificando uma investigação que considera também o nível estadual e, sobretudo, a realidade municipal de Rio Verde - GO, em que os impactos sociais e econômicos dos sinistros demandam respostas baseadas em evidências.

Apesar da urgência do problema, a produção científica dedicada aos sinistros de trânsito apresenta lacunas metodológicas importantes. Estudos que recorrem exclusivamente a modelos estatísticos clássicos conduzem, na maioria das vezes, a diagnósticos desagregados: analisam separadamente variáveis de infraestrutura, veículo ou condutor, sem capturar as inter-relações que atuam simultaneamente na ocorrência dos sinistros (CHANG; CHEN, 2005; LORD; MANNERING, 2010). Mesmo quando empregam modelos preditivos, tais pesquisas muitas vezes excedem os limites de validade externa, pois desconsideram a possibilidade de *data leakage,* isto é, a inclusão inadvertida de informações indisponíveis no momento da decisão, o que inflaciona artificialmente as métricas de desempenho (KAUFMAN *et al*., 2012; KAPOOR; NARAYANAN, 2023). Essa prática pode comprometer a generalização dos resultados e, portanto, a utilidade em cenários reais de gestão da segurança viária.

Adicionalmente, a literatura brasileira concentra grande parte de seus esforços em capitais ou regiões metropolitanas, enquanto municípios de porte médio, como Rio Verde, permanecem sub-representados.

Essas lacunas justificam a adoção de uma abordagem integrada, que combina análise exploratória, estatística espacial e *machine learning* com controle de vazamento de dados (*data leakage*). Tal combinação visa superar a fragmentação metodológica, oferecer diagnósticos espacialmente explícitos e produzir modelos preditivos realmente aplicáveis à realidade de Rio Verde – GO: município cuja hierarquia viária singular e dinâmica agroindustrial demandam soluções baseadas em evidência local.

Essas evidências dialogam diretamente com as metas do Plano Nacional de Redução de Mortes e Lesões no Trânsito (PNATRANS) e com o Plano Global da Década de Ação pela Segurança no Trânsito 2021-2030, ambos voltados a reduzir em pelo menos 50% as fatalidades até 2030 (BRASIL, 2023; ONU, 2020).

Este estudo contribui para a agenda ambiental, social e de governança (ESG) e para os Objetivos de Desenvolvimento Sustentável ao subsidiar informações que apoiem à proposição de intervenções que preservam vidas, reduzem custos hospitalares e otimizam a infraestrutura urbana, alinhando-se ao pilar social da ESG e ao ODS 3, que visa assegurar uma vida saudável e promover o bem-estar para todos (ONU, 2015). Ao incorporar técnicas de análise de dados e inovação tecnológica, também se alinha às metas do ODS 9, ao fomentar soluções baseadas em evidências voltadas à resiliência e sustentabilidade dos sistemas de transporte.

Essas contribuições reforçam o valor estratégico de estudos locais bem fundamentados: eles transformam dados rotineiros em inteligência territorial, conectam-se às diretrizes globais de desenvolvimento sustentável e oferecem rotas viáveis para que municípios de porte semelhante a Rio Verde adotem políticas de segurança viária baseadas em evidências.

A presente investigação considera exclusivamente registros de sinistros de trânsito sem vítimas, ocorridos na área urbana de Rio Verde (GO) entre 1.º de janeiro de 2021 e 31 de dezembro de 2024, conforme dados disponibilizados pela Agência Municipal de Mobilidade e Trânsito. A opção por esse recorte está relacionada aos registros acessíveis, sendo desejável, em estudos futuros, ampliar o recorte temporal e incluir sinistros com vítimas.

Em relação à metodologia, embora viabilize análises consistentes, apresenta vieses de registro que merecem consideração. Um dos mais recorrentes é a subnotificação, que ocorre quando apenas uma parcela dos eventos é formalmente reportada. Pesquisa como a de Alsop & Langley (2001) indica que sinistros mais graves, especialmente os que resultam em fatalidades, têm maior probabilidade de serem registrados, enquanto sinistros leves, que envolvem apenas danos materiais ou ferimentos menores, frequentemente não entram nas estatísticas oficiais.

Esse viés afeta a capacidade de avaliar corretamente os riscos e os fatores contribuintes para a acidentalidade, comprometendo o planejamento de estratégias preventivas. Além da subnotificação, a inconsistência e a incompletude dos registros são problemas recorrentes na coleta de dados de trânsito. Informações essenciais, como a severidade das lesões, as condições ambientais no momento do sinistro e o comportamento dos condutores, nem sempre são devidamente registrados. Em parte, pela falta de padronização nos procedimentos de coleta e uso de formulários inadequados, comprometendo a integração desses registros com outras bases de dados, como sistemas hospitalares e estatísticas de tráfego (Hauer & Hakkert, 1988).

Estudos nacionais e internacionais têm empregado bases de dados de acidentes de trânsito, inclusive aquelas compostas predominantemente por registros de sinistros sem vítimas, para compreender a distribuição espacial e temporal dos eventos e subsidiar políticas de segurança viária. No contexto brasileiro, Queiroz (2003) desenvolveu uma análise espacial dos acidentes em Fortaleza, utilizando dados georreferenciados para identificar locais críticos e padrões de concentração, ressaltando a relevância dessas informações mesmo em cenários com incompletude de registros. Em Santa Catarina,

Silva *et al*. (2020) investigaram a distribuição dos acidentes em rodovias estaduais a partir de dados administrativos, evidenciando que a análise espacial permite priorizar intervenções mesmo quando a base carece de detalhes sobre a gravidade.

No cenário internacional, trabalhos como o de Amoros, Martin e Laumon (2006) examinaram o viés de registro em bases oficiais, mostrando que sinistros sem vítimas fatais representam parcela significativa e fornecem informações úteis para identificar fatores de risco e áreas críticas. Hauer e Hakkert (1988) discutiram a importância da padronização na coleta de dados para assegurar comparabilidade entre estudos, enquanto Alsop e Langley (2001) demonstraram que a análise desses registros contribui para compreender padrões de comportamento no tráfego urbano.

Estudos recentes também reforçam a aplicabilidade de bases com registros não fatais. Moghadas *et al.* (2024) aplicaram modelos espaciais bayesianos para avaliar a influência de características viárias e socioeconômicas sobre a ocorrência de sinistros sem vítimas no Irã, evidenciando correlação com variáveis ambientais. Em Portugal, Pinho *et al.* (2023) utilizaram dados policiais de acidentes leves para desenvolver mapas de risco e prever áreas de maior probabilidade de ocorrência. Em países de alta renda, como Canadá e Nova Zelândia, pesquisas conduzidas por Levine *et al*. (1995) e Keall & Newstead (2020) mostraram que a inclusão de registros não fatais em análises espaciais aumenta a robustez estatística e a capacidade preditiva dos modelos.

No Brasil, estudos aplicados ao contexto urbano, como o de Lages *et al*. (2017), enfatizam que, embora haja subnotificação, bases contendo sinistros sem vítimas são insumos valiosos para diagnosticar problemas de engenharia de tráfego e planejar intervenções. De modo semelhante, trabalhos apresentados em congressos técnicos, como o de Camboriú (2019), destacam a utilidade de tais registros para análises de curto prazo e avaliação de medidas corretivas.

A literatura demonstra que, apesar das limitações inerentes, bases compostas por sinistros sem vítimas têm aplicação consolidada em pesquisas acadêmicas e em diagnósticos de segurança viária, oferecendo insumos estratégicos para identificação de *hotspots*, análise de fatores contribuintes e priorização de ações preventivas. No contexto deste estudo, *hotspot* refere-se a locais com alta concentração de sinistros de trânsito em determinado período, cuja identificação permite direcionar ações de engenharia, fiscalização e educação. Entre as metodologias consolidadas para a detecção, destacam-se a análise de densidade *kernel* (KDE), amplamente utilizada para estimar a intensidade espacial de ocorrências; o método empírico de Bayes, aplicado para ajustar estimativas

considerando variações aleatórias; e indicadores de risco derivados de taxas padronizadas por volume de tráfego ou população exposta (Anderson, 2009; Elvik, 2008). A metodologia proposta neste trabalho combina técnicas de estatística espacial com modelos de aprendizado de máquina, permitindo não apenas identificar áreas de maior concentração de sinistros sem vítimas, mas também explorar padrões temporais e variáveis associadas. Essa integração contribui para a análise e fornece subsídios mais precisos para o planejamento de intervenções preventivas no contexto urbano.

Para situar o recorte metodológico adotado neste estudo, no contexto mais amplo da acidentalidade no país, realizou-se uma análise comparativa com dados nacionais de sinistros com vítimas, considerando o período de 2021 a 2024 e indicadores consolidados pelo RENAEST/SENATRAN.

Os dados nacionais de sinistros com vítimas entre 2021 e 2024 apontam leve variação no volume total de ocorrências e nos indicadores de gravidade. Nesse período, o número de óbitos variou entre 23.745 (2021) e 21.525 (2024), enquanto a taxa de óbitos por 100 mil habitantes oscilou de 11,40 a 10,04. O percentual de óbitos por sinistro manteve-se entre 2,35% e 1,89%, refletindo mudanças nos padrões de severidade e na distribuição das ocorrências registradas.

Apesar da redução nas mortes, o volume total de sinistros manteve-se elevado, com picos em 2023 (1.165.899 registros) e pequena queda em 2024 (1.140.114). O número de feridos e ilesos permaneceu elevado, variando de 1,41 a 1,58 milhão por ano, reforçando a necessidade de ações preventivas contínuas.

Quando comparados com os sinistros sem vítimas, que constituem o foco deste estudo, esses dados permitem situar o recorte adotado em um contexto mais amplo, evidenciando que, embora as ocorrências fatais estejam em queda, a frequência de eventos menos graves permanece significativa e potencialmente indicativa de pontos críticos no espaço urbano. A análise comparativa, portanto, auxilia na compreensão da importância estratégica de monitorar e intervir também sobre esses registros, prevenindo a escalada para eventos de maior severidade.

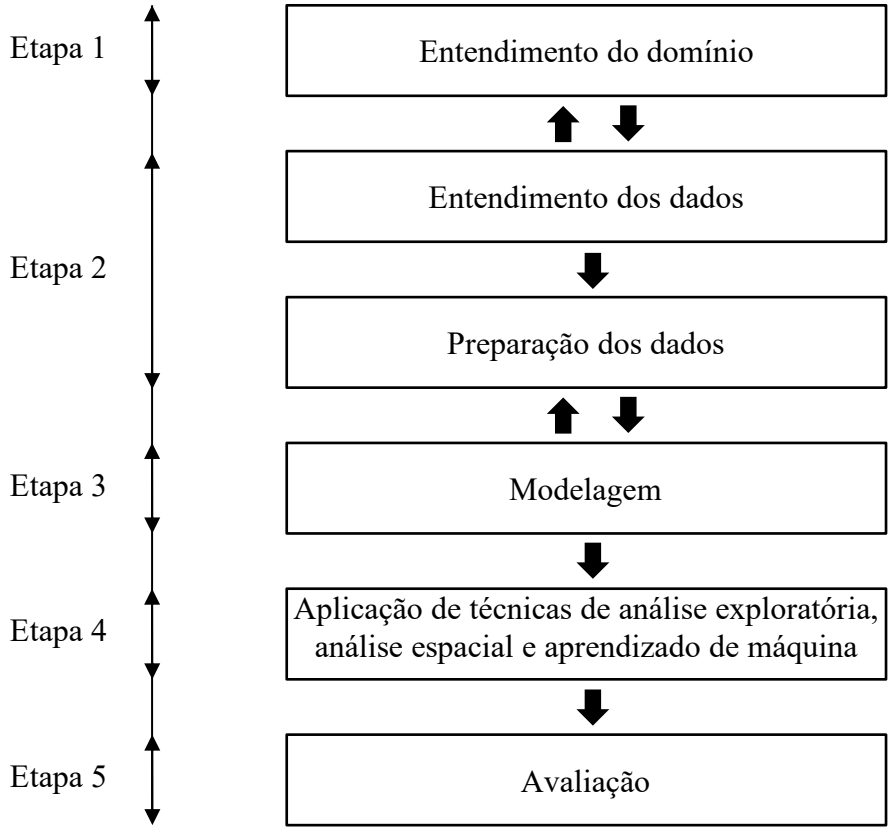
1.2. Metodologia

A metodologia desta pesquisa foi concebida para garantir que cada etapa, da coleta ao teste de modelos preditivos, responda de forma coerente às lacunas identificadas na

justificativa e alinhe-se aos três artigos que compõem a dissertação. O desenho adotado percorre um *pipeline* de pré-processamento estruturado e desdobra-se em abordagens estatística, espacial e de *machine learning*. Embora os conjuntos finais de dados utilizados apresentem pequenas variações de tamanho e abrangência temporal, especialmente no Capítulo III em relação aos Capítulos I e II, todos derivam de uma mesma fonte institucional, com critérios padronizados de tratamento e validação, assegurando que os resultados produzidos sejam metodologicamente compatíveis, com alto potencial de replicação e utilidade prática para gestores públicos e pesquisadores interessados em segurança viária baseada em evidências.

Na Figura 1 são apresentadas, em formato de fluxograma, as etapas metodológicas adotadas.

Figura 1- Método adotado na pesquisa.



Fonte de dados e contexto da pesquisa

O estudo fundamenta-se na base eletrônica de sinistros de trânsito mantida pela Agência Municipal de Mobilidade e Trânsito de Rio Verde (AMT), composta por boletins digitais, preenchidos por agentes de campo e armazenados em sistema próprio da

autarquia. Esses dados foram extraídos para planilha eletrônica e constituem a fonte primária de dados para os três capítulos da dissertação, com recortes temporais distintos, conforme a finalização de cada etapa analítica.

Nos Capítulos I e II, que tratam respectivamente da análise estatística descritiva e da análise espacial e espaciotemporal, foi utilizada a versão mais atualizada da base, compreendendo o período de 1.º de janeiro de 2021 a 31 de dezembro de 2024. Para o Capítulo III, dedicado à modelagem preditiva por meio de técnicas de *machine learning*, foi empregada a mesma base institucional, porém com recorte temporal entre 1.º de janeiro de 2021 e 31 de dezembro de 2023. Essa delimitação ocorreu porque na época da conclusão do terceiro artigo, a versão atualizada até 2024 ainda não estava disponível, impossibilitando a replicação imediata das análises com o conjunto mais amplo.

Ressalta-se, contudo, que os procedimentos metodológicos adotados no Capítulo III são plenamente reprodutíveis, o que viabiliza, em estudos futuros, a atualização das análises preditivas com a base unificada dos demais capítulos, ampliando a comparabilidade longitudinal dos resultados e fortalecendo a robustez dos modelos desenvolvidos.

Capítulo I

O primeiro artigo da dissertação tem como objetivo caracterizar, sob uma perspectiva descritiva, os sinistros de trânsito registrados em Rio Verde (GO) entre 2021 e 2024. A abordagem metodológica concentra-se na avaliação da completude da base, na estruturação do conjunto de variáveis temporais e na identificação de padrões gerais de ocorrência. Foram aplicadas estatísticas descritivas univariadas e bivariadas, análise de frequências relativas e absolutas, médias móveis, histogramas e matrizes de correlação. Também foram utilizados testes de independência para variáveis categóricas.

Capítulo II

O segundo artigo aprofunda a análise dos sinistros a partir de uma abordagem espacial e espaciotemporal. Foram aplicadas técnicas de geoprocessamento para mapear a concentração de ocorrências no território urbano de Rio Verde. Inicialmente, utilizou-se a Estimativa de Densidade por *Kernel* (KDE) para gerar superfícies contínuas de concentração de eventos, com grid fixo de 500 metros. Em seguida, aplicou-se o Índice de Moran Global para mensurar a autocorrelação espacial e os Indicadores Locais de Associação Espacial (LISA) para classificar áreas segundo a significância estatística e tipo de *cluster* (alto-alto, baixo-baixo etc.). Para análise em escala de segmento viário, com foco na Avenida Presidente Vargas, empregou-se o algoritmo DBSCAN, que

permite identificar agrupamentos densos ao longo de eixos lineares. O conjunto dessas técnicas viabiliza a identificação de zonas críticas e padrões de dispersão espacial dos sinistros.

Capítulo III

O terceiro artigo concentra-se na avaliação da viabilidade de modelos preditivos para variáveis relacionadas a sinistros, com ênfase no controle rigoroso de *data leakage*. Foram definidas onze variáveis-alvo, tanto categóricas quanto contínuas, modeladas por meio de algoritmos supervisionados e não supervisionados. Entre os algoritmos empregados destacam-se *Random Forest, Extra Trees, XGBoost, LightGBM* e Regressão Logística. O conjunto de preditores foi previamente tratado com técnicas de codificação, normalização e eliminação de atributos comprometidos por vazamento de dados.

Todas as etapas da pesquisa foram conduzidas em ambiente computacional controlado, com o objetivo de garantir rastreabilidade, transparência e possibilidade de reprodução integral dos resultados. O processamento e análise dos dados foram realizados em linguagem Python, utilizando bibliotecas para ciência de dados, tais como pandas, *numpy, matplotlib, seaborn, scikit-learn, xgboost, lightgbm, statsmodels, geopandas*, esda e PySAL. A criação de mapas foi conduzida com base em arquivos vetoriais (*shapefiles*) e camadas *raster* georreferenciadas.

Cabe destacar que esta pesquisa foi conduzida com observância aos princípios éticos da pesquisa científica, sendo utilizados única e exclusivamente para fins acadêmicos e de interesse público, respeitando os princípios da finalidade, necessidade e minimização previstos na LGPD. A utilização da base foi autorizada formalmente pela AMT, mediante solicitação institucional, e o estudo foi dispensado de submissão ao Comitê de Ética em Pesquisa por se tratar de análise de dados secundários públicos, sem qualquer tipo de intervenção direta com seres humanos.

1.3. Fundamentação teórica

1.3.1. Estatística descritiva e inferencial aplicada a sinistros

A avaliação de sinistros de trânsito inicia-se por estatísticas descritivas que quantificam frequência e gravidade. Contagens absolutas de ocorrências, feridos e óbitos

permitem caracterizar a magnitude do problema, ao passo que taxas por 100.000 habitantes ou por bilhão de veículos-quilômetro padronizam a exposição e viabilizam comparações interjurisdicionais (WASHINGTON; KARLAFTIS; MANNERING, 2020). Índices de gravidade como razão mortos-feridos complementam a análise, sintetizando a energia dissipada nos eventos e a carga sobre o sistema de saúde (ELVIK *et al.,* 2009).

Para variáveis categóricas, tabelas de contingência e testes de qui-quadrado avaliam independência entre natureza do sinistro, tipo de usuário da via e nível de lesão. Quando frequências esperadas são baixas, aplica-se o teste exato de Fisher, assegurando validade das inferências (MONTGOMERY; RUNGER, 2018). Em variáveis contínuas, por exemplo, idade das vítimas ou distância do centro urbano—normalidade é rara; logo, medidas de posição (mediana) e dispersão (intervalo interquartil) substituem média e desvio-padrão, e comparações utilizam testes não paramétricos de Mann-Whitney ou Kruskal-Wallis.

A modelagem inferencial dos números de sinistros exige regressão para contagens. O modelo de Poisson, fundamentado na suposição de média igual à variância, raramente atende aos dados empíricos, que exibem sobredispersão decorrente de heterogeneidade não observada entre segmentos viários. A regressão binomial negativa introduz parâmetro de dispersão para corrigir essa violação, melhorando a estimativa dos erros-padrão (CAMERON; TRIVEDI, 2013). Quando predominam zeros estruturais—vias sem registros no período—utilizam-se modelos inflados a zero ou mistos de Poisson-lognormal (HILBE, 2011).

A severidade do desfecho (ileso, ferido, morto) é variável politômica de ordem natural. A regressão logística ordinal, baseada na função logística cumulativa, estima a probabilidade acumulada de pertencer a categorias mais graves, controlando por idade, sexo, tipo de veículo e velocidade regulamentada do trecho. Caso o pressuposto de proporcionalidade dos *odds* seja violado, recorre-se ao modelo logístico multinomial (LORD; MANNERING, 2010).

Para inferências confiáveis, diagnostica-se multicolinearidade pelos fatores de inflação da variância, investiga-se influência de observações com a distância de Cook e testa-se a qualidade de ajuste por *deviance* residual e critério AIC. A incerteza final das estimativas é expressa em intervalos de confiança de 95% obtidos por reamostragem *bootstrap,* técnica robusta a distribuições desconhecidas (HAUER, 2001).

1.3.2. Análise temporal e espaciotemporal

A análise temporal de sinistros de trânsito fundamenta-se na construção de séries de contagem e em representações bidimensionais que combinam unidades temporais (hora, dia, mês, ano) para revelar padrões de variação sazonal, ciclos semanais e tendências de longo prazo. A suavização por médias móveis permite reduzir a variabilidade estocástica inerente a dados de baixa frequência e destacar oscilações estruturais relevantes (O'SULLIVAN; UNWIN, 2003).

A dimensão espacial é incorporada por estimativas de densidade *kernel* (KDE), técnica de alisamento que produz superfícies contínuas de intensidade a partir de coordenadas pontuais, possibilitando a identificação de áreas com maior concentração de eventos sem pressupor distribuição paramétrica prévia (SILVERMAN, 1986). A segmentação da base em recortes temporais gera sucessivas superfícies comparáveis, fornecendo indícios de persistência ou deslocamento de *hotspots* ao longo do tempo (XIE; YAN, 2013).

Para avaliar dependência espacial global, emprega-se o índice de autocorrelação de Moran, cujo resultado positivo indica agrupamento de altas ou baixas frequências, enquanto valores negativos sugerem dispersão. A decomposição local por Indicadores Locais de Associação Espacial (LISA) delimita unidades "*high-high*" ou "*low-low*" e permite monitorar a evolução espaciotemporal (ANSELIN, 1995; GETIS, 2007).

A clusterização baseada em densidade, executada com o algoritmo DBSCAN, identifica conglomerados de ocorrências contíguas em redes viárias e distingue ruído de padrões estruturados, sem impor número pré-definido de grupos. O método é adequado a distribuições lineares típicas de eixos rodoviários e mantém robustez frente a variação de forma e tamanho dos agrupamentos (ESTER *et al.*, 1996; CHAINEY; RATCLIFFE, 2013).

O emprego conjunto de séries temporais, KDE, autocorrelação espacial e clusterização densidade-baseada fornece arcabouço coerente para examinar simultaneamente quando e onde os sinistros ocorrem, subsidiando a priorização de intervenções em segurança viária, conforme dinâmica espaço-temporal demonstrada na literatura especializada.

1.3.3. Inteligência artificial

A Inteligência Artificial (IA) representa importante área na ciência da computação, em que fornece vasta gama de ferramentas para solucionar diversas formas de problema. Segundo Russell e Norvig (2013), no campo das ciências e engenharias, está presente e em escala crescente.

A definição do conceito de Inteligência Artificial, no entanto, apresenta desafios pela complexidade em determinar o significado do termo "inteligência".

A etimologia da palavra "inteligência" remonta ao latim *inter* e *legere*, significando "entre" e "escolher", respectivamente. Essa raiz etimológica leva à conclusão de que a inteligência é a capacidade de escolher entre alternativas. Por sua vez, "artificial" provém do latim *artificiale*, indicando algo não natural, criado ou construído pelos seres humanos. Russell e Norvig (2013) fornecem perspectiva ampla da IA descrevendo-a como um recurso capaz de automatizar e sistematizar tarefas complexas, destacando o valor em diversas esferas do conhecimento humano.

Ainda segundo Russell e Norvig (2013) quatro abordagens principais são delimitadas em IA. A primeira abordagem busca criar um "sistema que pensa como ser humano", envolvendo atividades como tomada de decisões, resolução de problemas e aprendizado. A segunda abordagem visa desenvolver um "sistema que atua como ser humano", criando máquinas para realizar funções que demandam inteligência humana. A terceira abordagem concentra-se em um "sistema que pensa racionalmente", utilizando modelos computacionais para estudar faculdades mentais. Por fim, a quarta abordagem trabalha na criação de um "sistema que atua racionalmente", focando no projeto de agentes inteligentes.

No âmbito da evolução da IA, ela desenvolveu-se em duas vertentes principais: abordagens centradas nos seres humanos e abordagens centradas na racionalidade. A primeira é uma ciência empírica que envolve hipóteses e confirmação experimental, enquanto a segunda emprega conceitos matemáticos em sua construção (Gomes, 2010).

O campo da IA evoluiu com a introdução de sistemas inteligentes, marcados por arquiteturas que incluíam base de conhecimentos, motor de inferência, módulo de explicação, módulo de aquisição de conhecimentos e interface com o usuário. As gerações seguintes contemplaram o aprendizado simbólico, influenciado por linguagens de programação como Prolog, Fortran, Cobol e Lisp. Métodos de aprendizado simbólico,

como Analogia, Instâncias, Evolução, Seleção, Reforço não Supervisionado, Bayesiano, Explicações e Indução, foram fundamentais na evolução da IA (Mitchell, 1997).

1.3.4. Machine Learning

O termo "*Machine Learning*", ou em português, "Aprendizado de Máquina", foi formalizado pelo engenheiro do MIT Arthur Samuel em 1959, sendo definido como uma disciplina que concede aos computadores a capacidade de adquirir conhecimento automaticamente por meio da experiência (Mitchell, 1997; Izbicki e dos Santos, 2020). Inserido no espectro da Inteligência Artificial (IA), é um campo de estudo que se baseia em algoritmos capazes de identificar padrões em conjuntos de dados, possibilitando a previsão de eventos futuros (Mitchell, 1997). O aprendizado ocorre sempre que há modificação na estrutura do programa ou na base de dados, sendo impulsionado por novos dados de entrada, resultando no aperfeiçoamento contínuo do desempenho. O avanço da capacidade de processamento e a disponibilidade de dados permitiram o desenvolvimento de modelos analíticos baseados no aprendizado de máquina.

Assim, o contexto contemporâneo é caracterizado pela crescente disponibilidade de dados em diferentes formatos, desencadeando um desafio na análise desses vastos conjuntos de informações. O foco central reside na extração de conhecimentos úteis para aprimorar processos decisórios (Izbicki e dos Santos, 2020).

O processo de *Machine Learning* oferece várias abordagens, sendo crucial entender o problema para escolher a metodologia apropriada. Segundo Mitchell (1997), destacam-se três métodos principais: Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado por Reforço.

No Aprendizado Supervisionado, treina-se manualmente uma rede neural com conjuntos de dados contendo entradas com vistas a reconhecer e internalizar os padrões que levam a uma determinada saída. O Aprendizado Não Supervisionado lida com dados abrangentes, processando-os para que a rede reconheça e agrupe informações naturalmente. Por fim, o Aprendizado por Reforço envolve a análise de padrões para aprender as melhores decisões, ajustando-se continuamente com base em erros identificados, atribuindo maior peso aos padrões de acertos no processamento subsequente. Estes três métodos são empregados para criar modelos analíticos a partir de

dados de treinamento, eliminando a necessidade de codificação explícita de regras (Russell; Norvig, 2013).

Além dos métodos destacados, o campo do *Machine Learning* oferece extensa variedade de algoritmos. Cada algoritmo é um conjunto de instruções e procedimentos matemáticos projetados para permitir que computadores aprendam padrões a partir dos dados disponibilizados. Visto que cada algoritmo apresenta suas vantagens e limitações, a escolha de um ou mais a serem aplicados para cada necessidade depende da sua natureza, e dos tipos de dados fornecidos. A melhor maneira de validar a eficácia do algoritmo é realizando de forma prática testes e validações.

Atualmente, a aplicação prática em ML tornou-se uma tarefa mais acessível, comparado às últimas décadas. A popularidade de linguagens de programação, tais como *Python*, que possui uma comunidade ativa de desenvolvedores, e aliado à sua sintaxe clara e as bibliotecas disponíveis para a comunidade, permite que pessoas que não são especialistas em programação possam de forma simplificada a implementação dos diversos tipos de algoritmos (Müller; Guido, 2016).

Das bibliotecas disponíveis, destaca-se a *Scikit-learn*, que permite o treinamento e a análise em diversos algoritmos de ML com a aplicação de poucas linhas de código.

2. OBJETIVOS

2.1. Objetivo geral

O objetivo geral desta pesquisa é analisar, na perspectiva exploratória e espacial, os sinistros de trânsito ocorridos entre 2021 e 2024 na área urbana de Rio Verde (GO).

2.2. Objetivos específicos

Os objetivos específicos consistem em:

a. Explorar os registros de sinistros de trânsito sem vítimas ocorridos em Rio Verde (GO) entre 2021 e 2024, com foco na identificação de padrões descritivos;

b. Analisar a distribuição espaço-temporal de sinistros de trânsito sem vítimas no município de Rio Verde (GO), entre 2021 e 2024, por meio de técnicas de análise espacial e agrupamento geográfico;
c. Avaliar a viabilidade de modelos preditivos confiáveis para sinistros de trânsito mediante a aplicação de metodologia de detecção e correção de *data leakage*.

3. CAPÍTULO I

ANÁLISE EXPLORATÓRIA DOS DADOS DE SINISTROS DE TRÂNSITO DE RIO VERDE - GOIÁS

RESUMO

Este estudo descreve os padrões espaço-temporais dos sinistros de trânsito sem vítimas registrados em Rio Verde-GO entre 2021 e 2024 e avalia a qualidade da base eletrônica da Agência Municipal de Mobilidade e Trânsito. Dos 7 926 registros originais, 7 911 apresentaram dados completos, após exclusão de inconsistências mínimas, demonstrando completude atribuída ao preenchimento digital padronizado. A distribuição anual manteve estabilidade: 1.834 sinistros em 2021, 2 031 em 2022, 1.962 em 2023 e 2.085 em 2024, com picos vespertinos em dias úteis. A concentração espacial abrangeu predominantemente os eixos centrais urbanos, especialmente Setor Central, Jardim Goiás e Bairro Popular, em que ocorrem 30,6% dos eventos. Colisões bilaterais constituíram 43,6% dos registros, abalroamentos 38,9% e choques em objeto fixo 14,2%. Automóveis responderam por 59,6% dos veículos envolvidos, seguidos por caminhonetes (17,2%) e motocicletas (5,8%). Homens de 18 a 35 anos compuseram a maioria dos participantes, e 84% dos sinistros ocorrem sob tempo bom e pista seca. Testes de alcoolemia foram aplicados em apenas 6,66% dos condutores, revelando positividade de 23% acima do limite legal. Esses achados indicam necessidade de intervenções direcionadas a interseções urbanas centrais e ampliação da fiscalização de álcool, enquanto confirmam o potencial analítico de bases eletrônicas completas para políticas públicas efetivas no município estudado.

Palavras-chave: sinistros de trânsito; análise espaço-temporal; Rio Verde; base de dados eletrônica; segurança viária.

ABSTRACT

This study describes the spatiotemporal patterns of traffic crashes recorded in Rio Verde, Goiás, Brazil, from 2021 to 2024 and evaluates the quality of the electronic database maintained by the Municipal Mobility and Traffic Agency. Of 7,926 original records, 7,911 contained complete data after minimal inconsistencies were removed, demonstrating completeness attributable to standardized digital entry. Annual distribution remained stable, with 1,834 crashes in 2021, 2,031 in 2022, 1,962 in 2023, and 2,085 in 2024, with afternoon peaks on working days. Spatial concentration lay mainly along central urban corridors, notably the Setor Central, Jardim Goiás, and Bairro Popular, which accounted for 30.6% of events. Bilateral collisions constituted 43.6% of records,

side-impact crashes 38.9%, and single-vehicle impacts with fixed objects 14.2%. Automobiles represented 59.6% of vehicles involved, followed by pickup trucks (17.2%) and motorcycles (5.8%). Male participants aged 18–35 years predominated, and 84% of crashes occurred under clear weather and dry pavement. Breath-alcohol tests were applied to only 6.66% of drivers, of whom 23% exceeded the legal limit. These findings indicate the need for targeted interventions at central urban intersections and strengthened alcohol enforcement, while confirming the analytical potential of complete electronic databases for effective public policy in the municipality studied.

Keywords: traffic crashes; spatiotemporal analysis; Rio Verde; electronic database; road safety.

3.1. Introdução

Sinistros ou sinistros de trânsito são definidos como eventos que resultam em danos a veículos, cargas, pessoas, animais ou ao meio ambiente, ocorrendo em vias terrestres ou áreas de circulação pública e envolvendo pelo menos um elemento em movimento. De acordo com o Departamento Nacional de Infraestrutura de Transportes (DNIT), essa definição abrange tanto prejuízos materiais quanto impactos no tráfego e na infraestrutura viária. A terminologia "sinistro de trânsito" tem sido adotada em substituição a "acidente de trânsito", conforme estabelecido pela Associação Brasileira de Normas Técnicas (ABNT) na revisão da norma NBR 10697. Essa mudança conceitual está alinhada ao Plano Nacional de Redução de Mortes e Lesões no Trânsito (PNATRANS) e à abordagem de Sistemas Seguros, reforçando a ideia de que tais eventos não são meramente fortuitos, mas podem ser prevenidos por meio de políticas e intervenções adequadas.

Esses eventos são caracterizados por sua imprevisibilidade e natureza multifatorial, envolvendo variáveis como veículos, comportamento humano, infraestrutura viária e condições ambientais. De acordo com a Organização Mundial da Saúde (OMS), os sinistros de trânsito constituem uma das principais causas de morte e lesões em escala global, configurando-se como um desafio significativo para a saúde pública e a mobilidade urbana (WHO, 2023).

Nas últimas décadas, os sinistros de trânsito tornaram-se um problema de grandes proporções em diversas partes do mundo, mesmo com o avanço nas tecnologias de segurança veicular e na implementação de políticas públicas voltadas à redução da acidentalidade. O aumento da frota de veículos, a expansão urbana desordenada e o crescimento populacional são fatores que contribuem para a complexidade desse fenômeno. Além das perdas humanas, os impactos econômicos são substanciais, abrangendo custos hospitalares, indenizações decorrentes de invalidez, morte ou danos corporais, danos materiais e prejuízos à produtividade, onerando significativamente os sistemas de saúde e seguridade social (BACCHIERI; BARROS, 2011).

Ainda segundo Bacchieri e Barros (2011), estes eventos apresentam diversidade de causas, podendo estar relacionados a fatores humanos, como imprudência, fadiga, uso de substâncias psicoativas e desatenção; fatores ambientais, incluindo condições climáticas adversas e falhas na sinalização; e fatores mecânicos, como falhas nos sistemas

de frenagem e pneus em más condições. A análise aprofundada das causas e dinâmicas desses eventos é essencial para a formulação de políticas preventivas eficazes.

No Brasil, os sinistros de trânsito continuam sendo uma das principais causas de mortalidade, representando um desafio significativo para a saúde pública e a segurança viária. Dados do Ministério da Saúde de 2023 apontam que o país registrou 33.743 mortes no trânsito, um número próximo ao de 2022, quando foram contabilizados 33.894 óbitos, evidenciando a persistência do problema apesar das medidas de fiscalização e educação viária implementadas nos últimos anos O impacto desses eventos não se restringe às vítimas e seus familiares, estendendo-se para a sociedade como um todo, com elevados custos hospitalares, previdenciários e produtivos.

Além da elevada taxa de letalidade, os sinistros de trânsito no Brasil ocorrem de forma heterogênea, sendo influenciados por fatores como a densidade populacional, a infraestrutura viária e o comportamento dos condutores. A análise dos perfis das vítimas indica que a faixa etária entre 20 e 29 anos é uma das mais afetadas, sugerindo uma vulnerabilidade específica desse grupo no trânsito, segundo dados da Confederação Nacional dos Municípios. Paralelamente, o aumento da frota de veículos e a urbanização desordenada contribuem para a complexidade do problema, tornando essencial a formulação de políticas públicas baseadas em dados concretos e confiáveis (CNM, 2012).

Diante da persistência dos altos índices de acidentalidade no Brasil, torna-se imprescindível compreender não apenas as causas desses eventos, mas também a maneira como são registrados e analisados. A precisão e a confiabilidade dos dados coletados são determinantes para a formulação de políticas eficazes e para a redução do número de sinistros.

Nesse sentido, a coleta de dados sobre sinistros de trânsito desempenha papel central na formulação de políticas públicas eficazes voltadas para a segurança viária. A análise desses dados permite que gestores e pesquisadores desenvolvam estratégias baseadas em evidências para reduzir o número de sinistros e as consequências. De acordo com a Organização das Nações Unidas (ONU, 2020), a "Década de Ação para a Segurança no Trânsito 2021-2030" tem como um dos pilares a melhoria na coleta e na análise de dados, enfatizando que a disponibilidade de informações confiáveis é um requisito fundamental para monitorar o progresso das intervenções e avaliar a efetividade (ONU, 2021). Sem um banco de dados preciso, torna-se complexo identificar padrões de ocorrência, diagnosticar fatores de risco e projetar medidas preventivas adequadas.

A forma como os dados são registrados influencia diretamente na qualidade das análises realizadas. O processo de coleta pode ocorrer por meio de relatos das vítimas, registros policiais ou informações de instituições responsáveis pelo monitoramento viário. Entretanto, a existência de erros nesses registros, seja por imprecisão na descrição dos fatos ou pela ausência de variáveis essenciais, compromete a identificação de áreas críticas e reduz a confiabilidade das avaliações sobre a segurança viária (Hauer & Hakkert, 1988). Consequentemente, decisões baseadas em dados inconsistentes podem direcionar recursos de forma inadequada, impactando negativamente as estratégias de mitigação de riscos no trânsito.

Estudos como os de Mello-Jorge (1990) já destacavam a importância da qualidade da informação sobre sinistros para o planejamento adequado de políticas públicas. Variáveis como a localização exata, a condição da via, o perfil dos envolvidos e as circunstâncias do sinistro são fundamentais para que as análises sejam precisas e contribuam para a implementação de medidas preventivas eficazes. A coleta estruturada e confiável de dados sobre sinistros de trânsito deve ser considerada um dos pilares da segurança viária. O desenvolvimento de medidas de prevenção depende, em grande parte, da exatidão das informações disponíveis, visto que inconsistências podem comprometer a tomada de decisões. Dessa forma, aprimorar os sistemas de registro e garantir a fidedignidade dos dados são desafios fundamentais para reduzir os impactos e promover a mobilidade segura nas vias públicas.

Um dos problemas mais recorrentes é a subnotificação, que ocorre quando apenas uma parcela dos eventos é formalmente reportada. Pesquisas indicam que sinistros mais graves, especialmente os que resultam em fatalidades, têm maior probabilidade de serem registrados, enquanto sinistros leves, que envolvem apenas danos materiais ou ferimentos menores, frequentemente não entram nas estatísticas oficiais (Alsop & Langley, 2001). Esse viés afeta a capacidade de avaliar corretamente os riscos e os fatores contribuintes para a acidentalidade, comprometendo o planejamento de estratégias preventivas. Além da subnotificação, a inconsistência e a incompletude dos registros são problemas recorrentes na coleta de dados de trânsito. Informações essenciais, como a severidade das lesões, as condições ambientais no momento do sinistro e o comportamento dos condutores, nem sempre são devidamente registradas. Isso ocorre, em parte, pela falta de padronização nos procedimentos de coleta e uso de formulários inadequados, que compromete a integração desses registros com outras bases de dados, como sistemas hospitalares e estatísticas de tráfego (Hauer & Hakkert, 1988).

Diante dos desafios relacionados à acidentalidade viária e à qualidade dos dados coletados, torna-se fundamental compreender a realidade local para subsidiar políticas públicas mais eficazes. Nesse contexto, este estudo propõe uma análise exploratória dos sinistros de trânsito ocorridos no município de Rio Verde - GO entre os anos de 2021 e 2024, com foco no entendimento e na identificação de padrões relevantes.

3.2. Material e Métodos

A técnica de análise exploratória de dados é empregada para identificar padrões, tendências e correlações nos dados coletados. Esta abordagem, conforme descrito por Levine *et al*. (1996), envolve a coleta, caracterização e apresentação de dados para descrever de forma abrangente as características observadas, facilitando a visualização de complexidades e nuances dos dados.

Conforme Bussab e Morettin (2005), a estatística oferece uma variedade de ferramentas descritivas, como gráficos, tabelas e índices, que são essenciais para a organização e síntese dos dados, bem como reforça a importância dos gráficos como representações visuais que elucidam a evolução dos fenômenos ou as relações entre variáveis envolvidas, oferecendo compreensão clara e imediata dos padrões de sinistros.

Este estudo adota uma abordagem descritiva de corte transversal, caracterizada pela observação de um conjunto de dados em período específico, utilizando dados secundários sobre sinistros de trânsito na malha urbana do município de Rio Verde, Goiás, ao longo do período de 2021 a 2024.

Coleta de dados

A coleta de dados deste estudo foi realizada através de duas fontes. A primeira fonte foi encaminhada pelo Núcleo Integrado de Análise Criminal e Inteligência, da Superintendência Integrada de Tecnologia em Segurança Pública de Rio Verde, que encaminhou o Relatório de Análise Criminal nº 109/2023 contendo dados de ocorrência mensal, óbitos registrados em sinistros, sexo, faixa etária, condições da via e bairros com a maior incidência. O relatório evidenciou que o processamento dos atendimentos das instituições Corpo de Bombeiros Militar do Estado de Goiás e Polícia Militar de Goiás, apesar de serem integrados pelo Registro de Atendimento Integrado – RAI, possui certa divergência, pois eventualmente o atendimento pré-hospitalar pode ser realizado pelo

Serviço de Atendimento Móvel – SAMU, que não possui acesso ao banco de dados. O relatório ressalta ainda que existem sinistros de trânsito que podem ser registrados pela Agência Municipal de Mobilidade e Trânsito – AMT, que também não está integrada ao sistema RAI, e o Núcleo Integrado não possuí acesso ao banco de dados.

Outra base de dados, fornecida pela Agência Municipal de Mobilidade e Trânsito de Rio Verde (AMT), contemplava 7.926 registros de sinistros ocorridos entre 01 de janeiro de 2021 a 31 de dezembro de 2024. As informações que caracterizavam cada sinistro incluíam: data, horário, endereço, número de veículos envolvidos, tipo de sinistro e causa provável.

Dos 7.926 registros disponibilizados, identificou-se que 15 não continham o link de acesso ao formulário do boletim de sinistro de trânsito. Em função da ausência dessa informação fundamental para a análise dos atributos associados a cada ocorrência, tais registros foram excluídos da base de dados. Assim, a base final analisada passou a conter 7.911 registros válidos.

Pré-processamento de dados

Ao analisar os valores incorretos e inexistentes nos registros de sinistros de trânsito das duas bases de dados, observa-se que, embora a maioria destes dados coletados esteja devidamente registrada, há inconsistências pontuais que podem impactar análises mais detalhadas. A identificação desses valores ausentes ou incorretos é essencial para garantir a qualidade da base de dados e aprimorar futuras coletas, possibilitando maior precisão na identificação de padrões de acidentalidade.

Para extrair e estruturar informações detalhadas dos registros de sinistros, foi desenvolvido um processo automatizado em três etapas, utilizando a linguagem *Python* e bibliotecas como pandas, *requests* e *BeautifulSoup*. Primeiro, a base original em formato CSV foi carregada, contendo links para os boletins de ocorrência de sinistro, disponíveis online. Com função automatizada, o conteúdo dessas páginas foi acessado e extraído em texto limpo, sendo salvo como arquivos .txt, identificados por um código único.

Na segunda etapa, os arquivos de texto foram lidos e processados para separar as seções dos relatórios, que foram transformadas em pares chave-valor.

Os conteúdos extraídos foram processados para identificar e estruturar apenas as variáveis operacionais relacionadas ao evento. Qualquer campo que pudesse identificar indivíduos foi descartado antes da consolidação da base analítica.

Por fim, os dados foram padronizados, reestruturados e integrados. Os registros de envolvidos foram convertidos para o formato “largo”, permitindo que cada linha representasse todos os envolvidos de um mesmo boletim. Em seguida, esses dados foram combinados com os dados gerais dos sinistros, gerando um único conjunto consolidado e livre de duplicações. O resultado foi exportado em um arquivo Excel, pronto para ser utilizado em análises estatísticas, geográficas e operacionais.

O tratamento dos dados após a extração foi executado utilizando o *software* Microsoft Excel. As principais ações incluíram a correção de formatos de data e hora, a padronização de entradas textuais em campos de categorias de sinistros, condições meteorológicas e tipos de veículos envolvidos, bem como a verificação e correção de desvios ou dados duplicados. Além disso, foram identificadas e corrigidas inconsistências, tais como entradas ilógicas ou contraditórias, como por exemplo, registros marcados simultaneamente com condições de tempo seco e chuvoso.

O processamento automático foi submetido a procedimento de validação por verificação manual. Para isso, selecionou-se uma amostra aleatória de boletins de ocorrência em formato original (HTML) e comparou-se o conteúdo com os registros estruturados, após a raspagem e integração. Essa conferência assegurou que os campos extraídos correspondessem fielmente às informações originais.

A partir da contagem das variáveis "data de nascimento" e "sexo" associadas a cada boletim de ocorrência, foi possível inferir a quantidade de indivíduos envolvidos por registro, permitindo a construção de uma variável derivada que representa o total de envolvidos por sinistro.

O Quadro 1 apresenta a relação de variáveis utilizadas para este estudo, com a descrição e os respectivos valores, após a extração dos dados dos formulários e o pré-processamento.

Quadro 1 - Relação de variáveis e a respectiva descrição.

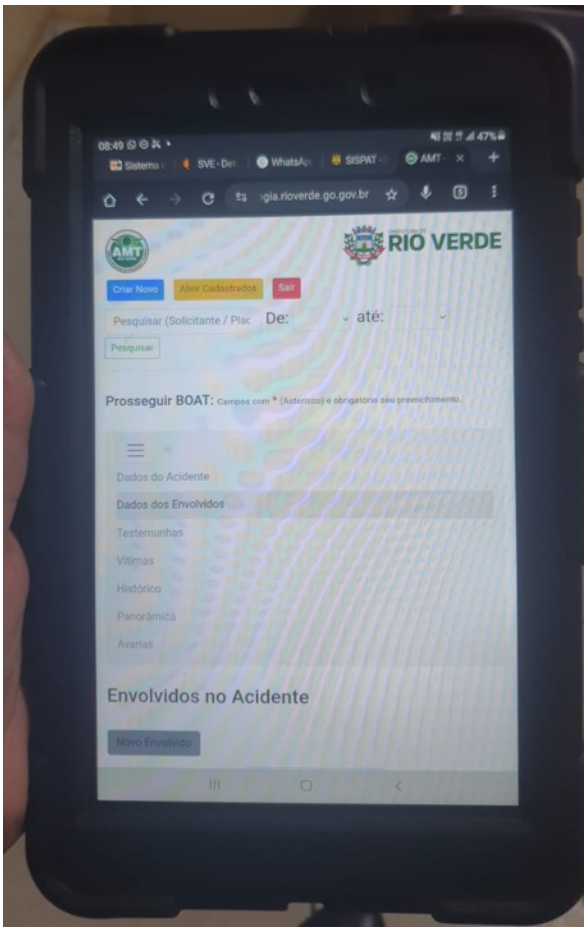
Variável	Descrição
Boletim	Número do boletim de ocorrência
Data	Data do sinistro
Mês	Mês em que ocorreu o sinistro
Ano	Ano em que ocorreu o sinistro
Dia semana	Dia da semana em que ocorreu o sinistro
Hora	Horário do sinistro
Período	Período do dia em que ocorreu o sinistro
Coordenada	Coordenada geográfica do sinistro
Bairro	Nome do bairro em que ocorreu o sinistro
Natureza	Tipo de sinistro ocorrido
Controle tráfego	Indicação de tráfego no local
Zona	Classificação da zona em que ocorreu o sinistro
Pista	Descrição das características da pista em que ocorreu o sinistro
Pavimento	Tipo de pavimento da via
Condições da via	Descrição da via em que ocorreu o sinistro
Condições do tempo	Condições climáticas no momento do sinistro
Envolvidos	Quantidade de envolvidos
Data nascimento	Data de nascimento do(s) envolvido(s)
Sexo	Sexo dos(s) envolvido(s)
Exame	Realização do exame toxicológico
Dosagem	Valor do exame toxicológico
Tipo	Tipo de veículo do(s) envolvido(s)

A qualidade dos dados é um aspecto fundamental para a confiabilidade das análises e a tomada de decisões baseadas em evidências. Segundo Mahanti (2019), “a qualidade dos dados é a capacidade dos dados de satisfazer os requisitos técnicos, de sistema e de negócios declarados de uma empresa”. No contexto da ciência de dados aplicada, garantir a integridade, completude e consistência da base é essencial para assegurar a robustez dos resultados e a validade das inferências.

Neste estudo, destaca-se que os dados foram coletados por meio de formulários eletrônicos, utilizados pela Agência Municipal de Mobilidade e Trânsito (AMT) de Rio Verde. Esse processo automatizado de entrada de dados minimiza significativamente o risco de erros de preenchimento e inconsistências manuais, comuns em registros feitos em papel ou por meio de digitação não supervisionada. Como resultado, a necessidade de pré-processamento foi bastante reduzida. As etapas de tratamento limitaram-se, em grande parte, à padronização de formatos, como a conversão de variáveis temporais (por exemplo, horário dos sinistros) para estruturas compatíveis com os modelos de análise.

A Figura 2 ilustra a interface do formulário eletrônico utilizado pela AMT de Rio Verde, e evidencia o controle automatizado de campos e a estruturação dos dados em conformidade com os requisitos técnicos de entrada. Esse padrão contribui decisivamente para a elevação da qualidade dos dados e, consequentemente, para a confiabilidade das análises realizadas.

Figura 2 – Formulário eletrônico de coleta de dados de sinistro.



A seguir, procede-se à apresentação das análises e resultados obtidos, com os dados tratados, com vistas a explorar de forma sistemática os dados.

3.3. Resultados e Discussão

A análise descritiva permite compreender melhor os padrões e tendências apresentados pelos dados, fornecendo uma base sólida para interpretação e tomada de decisão. Por meio de consultas estruturadas, agrupamentos e ordenações, é possível gerar

visualizações gráficas e extrair informações relevantes que auxiliaram as próximas etapas da análise. Segundo Yamamoto (2009), análise descritiva refere-se ao conjunto de técnicas utilizadas para organizar, resumir e descrever os aspectos principais de um conjunto de dados, fornecendo uma visão geral das características dos dados sem tirar conclusões além dos dados analisados. Os resultados da distribuição dos sinistros ao longo do tempo auxiliam na compreensão de padrões sazonais e horários. Os dados analisados foram organizados de acordo com o ano, o mês, o dia da semana e o período do dia, permitindo uma avaliação da frequência dos eventos em diferentes escalas temporais.

Análise temporal

Entre os anos de 2021 e 2024, foram registrados, respectivamente, 1.833, 2.031, 1.962 e 2.085 sinistros de trânsito no município de Rio Verde. Esses valores correspondem, na mesma ordem, a 23,17%, 25,67%, 24,80% e 26,36% do total de ocorrências registradas no período.

No recorte mensal, janeiro apresentou 595 sinistros (7,52%), fevereiro 602 (7,61%), março 703 (8,89%), abril 649 (8,20%), maio 719 (9,09%), junho 676 (8,55%), julho 674 (8,52%), agosto 724 (9,15%), setembro 656 (8,29%), outubro 629 (7,95%), novembro 680 (8,60%) e dezembro 604 (7,63%).

Na análise por dia da semana, com base na soma dos registros entre 2021 e 2024, domingo concentrou 650 ocorrências (8,22%), segunda-feira 1.253 (15,84%), terça-feira 1.255 (15,86%), quarta-feira 1.243 (15,71%), quinta-feira 1.227 (15,51%), sexta-feira 1.260 (15,93%) e sábado 1.023 (12,93%).

No recorte por turno, considerando a soma dos registros no período analisado, a tarde concentrou o maior volume de sinistros, com 3.296 ocorrências (41,66%), seguida da manhã, com 2.486 (31,42%), e da noite, com 1.898 (23,99%). O turno da madrugada apresentou o menor número de registros, totalizando 231 sinistros, o que corresponde a 2,92% do total.

Os mapas de calor oferecem uma representação visual da distribuição temporal dos sinistros de trânsito. A partir dessa perspectiva, a Figura 3 apresenta a concentração de ocorrências por mês ao longo dos quatro anos analisados, enquanto a Figura 4 evidencia a distribuição por dia da semana e a Figura 5 organiza os registros segundo os turnos do dia.

Figura 3 - Mapa de calor de sinistros por mês e ano.

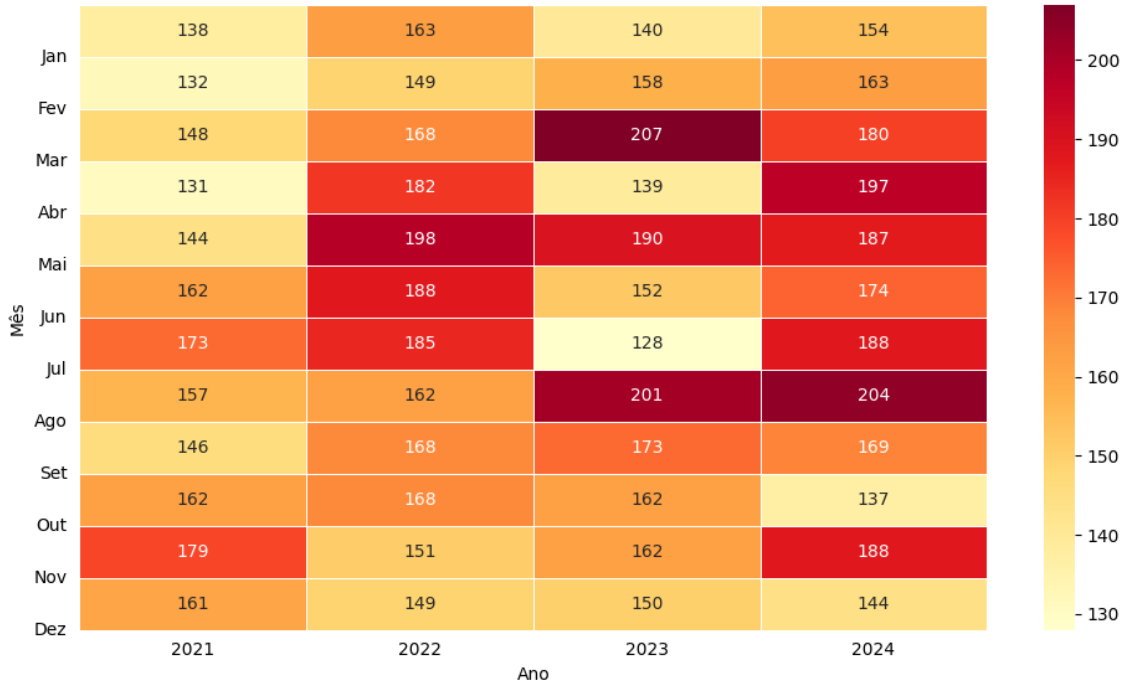


Figura 4 - Mapa de calor de sinistros por dia da semana e ano.

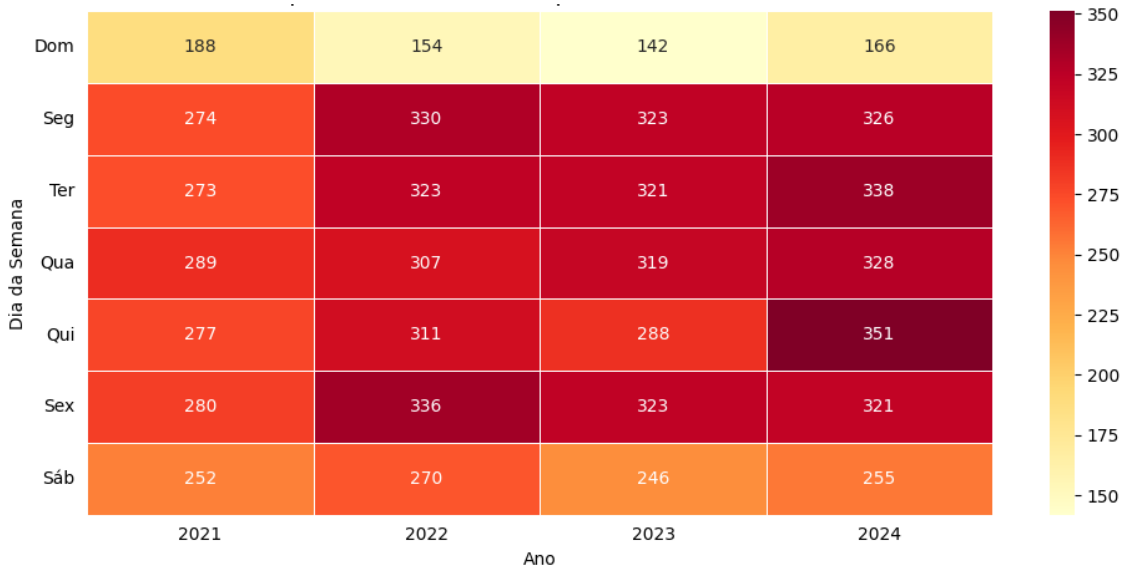
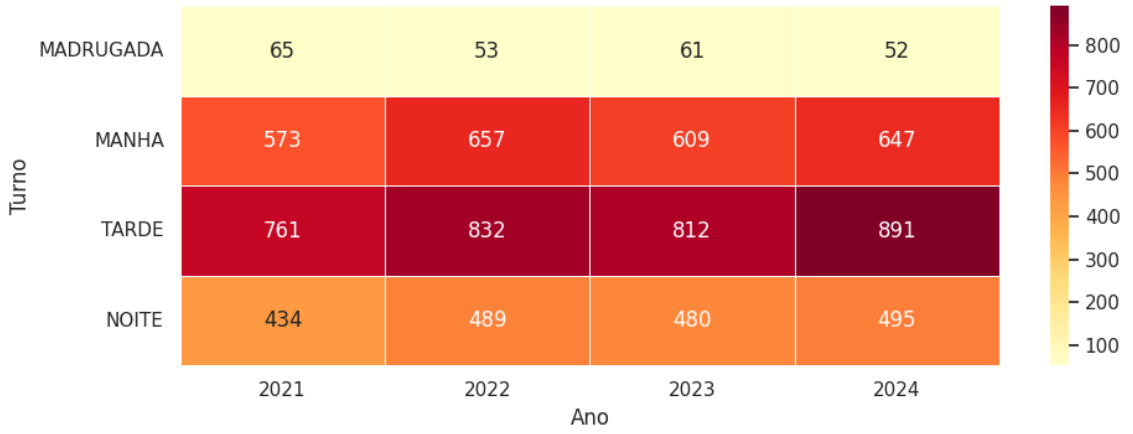


Figura 5 - Mapa de calor de sinistros por turno e ano.



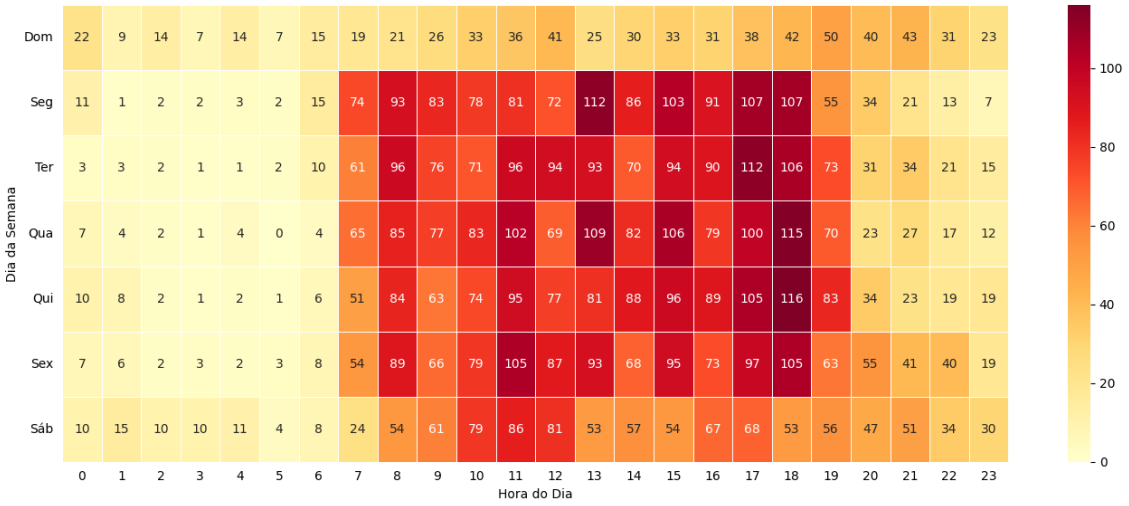
A quantidade de sinistros diários ao longo de um ano, comparando os períodos de 2021 a 2024, é evidenciada na Figura 6, com a aplicação da média móvel de 30 dias.

Observa-se que a frequência diária de sinistros apresenta grande variabilidade, com oscilações frequentes ao longo do ano. Esse comportamento sugere que os sinistros não ocorrem de maneira uniforme, mas sim influenciados por fatores específicos de cada período. Picos de sinistros podem ser percebidos em diferentes momentos do ano, o que pode estar relacionado a datas comemorativas, períodos de maior movimentação urbana e variações nas condições meteorológicas.

Figura 6 – Média móvel (30 dias) do número de sinistros por dia.

A Figura 7 relaciona a distribuição dos sinistros ao longo dos dias da semana e a hora do dia, destacando os períodos de maior concentração de sinistros com base na totalização dos dados entre 2021 e 2024.

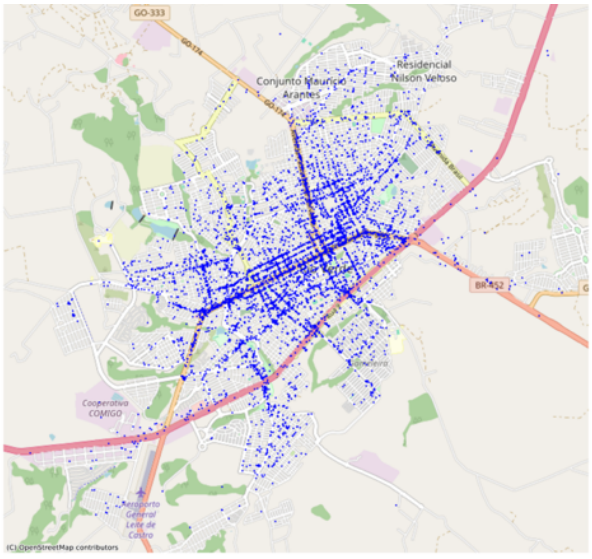
Figura 7 - Mapa de calor de sinistros por dia da semana e hora.



Análise espacial

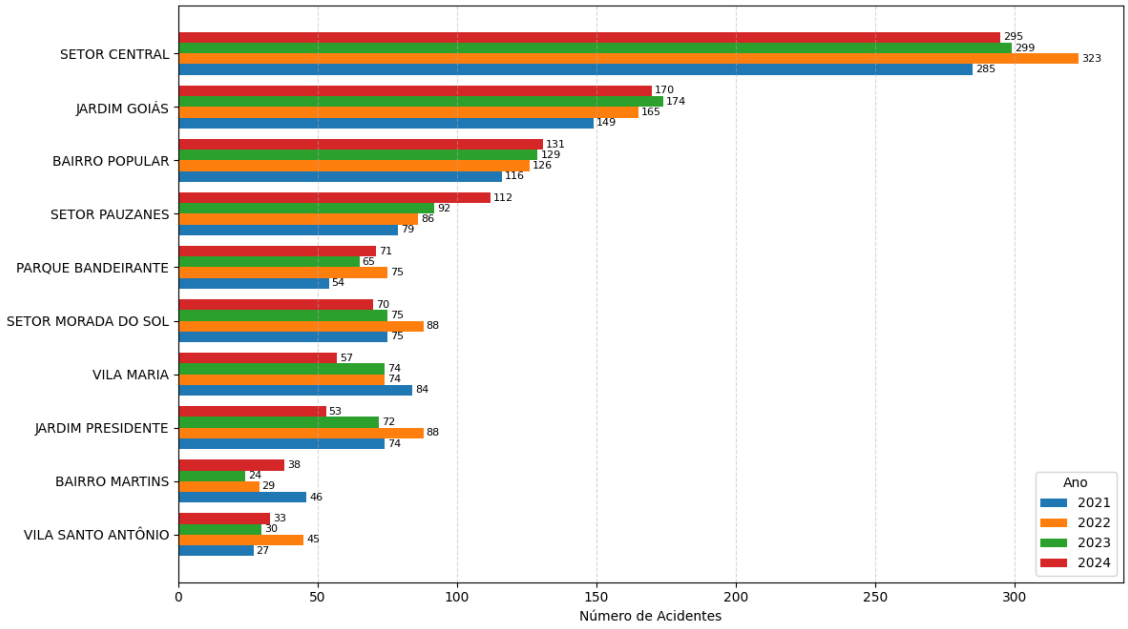
A Figura 8 apresenta a distribuição espacial dos sinistros de trânsito no município de Rio Verde entre os anos de 2021 e 2024. A visualização, construída a partir da dispersão dos pontos de ocorrência georreferenciados, evidencia uma concentração significativa de sinistros na zona urbana central do município, com dispersões secundárias em regiões periféricas e ao longo dos principais corredores viários. A utilização de pontos com baixa opacidade permite identificar áreas com sobreposição elevada, as quais se destacam visualmente como zonas de maior densidade de registros.

Figura 8 – Distribuição espacial dos sinistros em Rio Verde.



A Figura 9 apresenta a distribuição anual dos sinistros de trânsito nos dez bairros com maior número de registros em Rio Verde. Os três bairros com maior incidência, Setor Central (15,9%), Jardim Goiás (8,32%) e Bairro Popular (6,35%), concentram, juntos 30,57% de todos os registros da base. No total, a base contempla registros em 147 bairros distintos, reforçando a ampla dispersão espacial das ocorrências no município.

Figura 9 – Top 10 bairros com maior número de sinistros.



Entre os anos de 2021 e 2024, alguns bairros de Rio Verde apresentaram variações significativas no número de sinistros de trânsito. Do ponto de vista do crescimento percentual, destacam-se bairros que, embora tenham iniciado o período com um número reduzido de ocorrências, registraram aumentos expressivos. É o caso do Jardim Mondale e do Jardim Brasília, que passaram de 1 sinistro em 2021 para 8 em 2024, correspondendo a um crescimento de 700% em ambos. O Residencial Jardim Bougainville teve aumento de 3 para 17 ocorrências no mesmo intervalo, resultando em elevação de 466,7%. Outros bairros com crescimento relevante incluem a Vila Menezes (240%), o Jardim Adriana (233,3%), o Residencial Atalaia (300%), e o Condomínio Nova Aliança Premium (200%).

Por outro lado, alguns bairros apresentaram decréscimos acentuados, com destaque para aqueles que, apesar de registrarem sinistros em 2021, zeraram as

ocorrências em 2024. É o caso da Vila Mariana Prolongamento I e II, Vila Santo André, Distrito Agroindustrial de Rio Verde I (DARV I), Vila Amália II e outros, todos com variação negativa de 100%.

Análise das características e condições dos sinistros

A Tabela 1 apresenta a distribuição dos sinistros de trânsito por tipologia, considerando os registros entre os anos de 2021 e 2024. Observa-se forte concentração em três categorias principais: colisão (44%), abalroamento (39%) e choque em objeto fixo (14%), que, somadas, correspondem a 97% do total de ocorrências. Essa predominância sugere que a maioria dos sinistros envolve veículos em movimento, possivelmente em cruzamentos, mudanças de faixa ou situações de tráfego intenso em áreas urbanas.

Tabela 1 – Distribuição dos sinistros por natureza.

Natureza	2021		2022		2023		2024		Total	
Colisão	781	42,6%	872	42,9%	881	44,9%	918	44,0%	3.452	43,6%
Abalroamento	729	39,8%	779	38,4%	746	38,0%	826	39,6%	3.080	38,9%
Choque em objeto fixo	250	13,6%	324	16,0%	272	13,9%	276	13,2%	1.122	14,2%
Outro	45	2,5%	35	1,7%	47	2,4%	50	2,4%	177	2,2%
Atropelamento	21	1,1%	15	0,7%	7	0,4%	9	0,4%	52	0,7%
Tombamento	4	0,2%	2	0,1%	3	0,2%	5	0,2%	14	0,2%
Capotamento	3	0,2%	4	0,2%	3	0,2%	0	0%	10	0,1%
Atropelamento animal	0	0%	0	0%	3	0,2%	1	0,05%	4	0,1%
Total	**1.833**	**100%**	**2031**	**100%**	**1.962**	**100%**	**2.085**	**100%**	**7911**	**100%**

A análise temporal revela que a distribuição percentual dessas categorias manteve-se relativamente estável ao longo dos quatro anos, indicando que os fatores estruturais associados à dinâmica do trânsito local, como o traçado viário, sinalização e comportamento dos condutores, pouco alteraram no período.

Outras tipologias, como atropelamento (0,7%), tombamento (0,2%), capotamento (0,1%) e atropelamento de animal (0,1%), apresentam incidência bastante reduzida.

A categoria “Outro” concentra 2,2% dos registros, valor superior ao de algumas tipologias específicas, podendo indicar falhas na padronização do registro ou uso excessivo dessa classificação em casos pouco detalhados.

A análise da distribuição dos veículos envolvidos em sinistros, considerando os cinco tipos mais frequentes, revela a predominância expressiva dos automóveis, que somam 9.051 registros, correspondendo a 59,64% do total consolidado. Em seguida, destacam-se as caminhonetes com 2.607 registros (17,17%), as motocicletas com 873 registros (5,75%), os caminhões com 670 registros (4,42%) e as camionetas com 540 registros (3,56%). Os demais tipos de veículos, totalizam 1.373 registros, representando 9,05% dos casos.

Ao cruzar essas informações com a natureza dos sinistros (Tabela 2), observa-se que a maior concentração de automóveis ocorre em eventos classificados como colisão (4.328) e abalroamento (3.454), que juntos respondem por mais de 86% dos envolvimentos com esse tipo de veículo. A distribuição das caminhonetes segue padrão semelhante, com forte presença nas colisões (1.257) e abalroamentos (982). As motocicletas, por sua vez, mantêm presença significativa tanto em abalroamentos (435) quanto em colisões (340).

Os caminhões aparecem com maior frequência em abalroamentos (255) e colisões (218). Já as camionetas, embora em número inferior, apresentam padrão semelhante aos veículos de maior porte, com maior ocorrência em colisões e abalroamentos.

Tabela 2 – Distribuição dos sinistros por natureza e tipo de veículo.

Natureza	**Automóvel**		**Caminhão**		**Camin-honete**		**Camioneta**		**Motocicleta**		**Outros veículos**	
Colisão	4.328	47,8%	218	32,5%	1.257	48,2%	239	44,3%	340	38,9%	515	37,5%
Abalroa-mento	3.454	38,2%	255	38,1%	982	37,7%	213	39,4%	435	49,8%	549	40,0%
Choque em objeto fixo	1.004	11,1%	168	25,1%	299	11,5%	71	13,1%	70	8,0%	267	19,4%
Outras naturezas	265	2,9%	29	4,3%	69	2,6%	17	3,1%	28	3,2%	42	3,1%
Total	**9.051**	**100%**	**670**	**100%**	**2.607**	**100%**	**540**	**100%**	**873**	**100%**	**1.373**	**100%**

No que se refere à condição do tempo (Figura 10), observa-se que a maioria dos sinistros ocorreu sob tempo bom (6.646), representando um volume significativamente superior em relação às demais categorias. Mesmo assim, nota-se a existência de um número relevante de sinistros em períodos nublados (678) e chuvosos (410), o que totaliza 1.088 (13,75%). Quanto à condição da via (Figura 11), verifica-se que a maioria dos sinistros ocorreu em vias secas, o que acompanha a distribuição observada no gráfico da condição do tempo. Os registros evidenciam também que a maior parte dos sinistros

ocorreu em vias asfaltadas, representando 98,51%, seguido de concreto, com 0,68% e paralelepípedo com 0,51%. As demais categorias somadas representam 0,3% dos registros (terra e cascalho). Os dados apresentam ainda que 99,22% dos sinistros ocorrerem em área urbana, enquanto apenas 0,78% foram registrados em zona rural.

Figura 10 - Sinistros por condição do tempo.

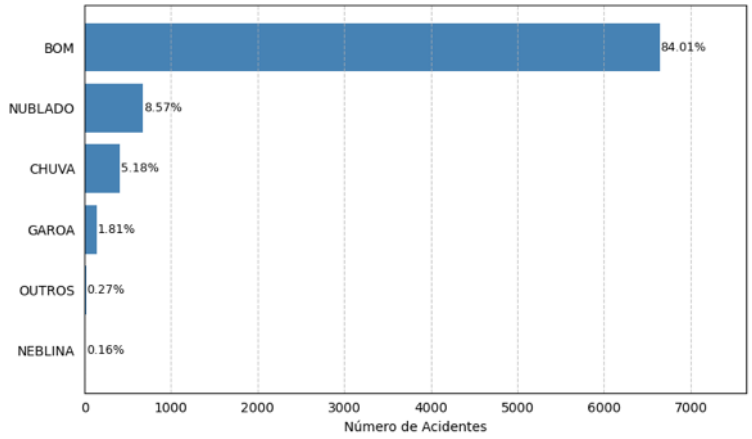
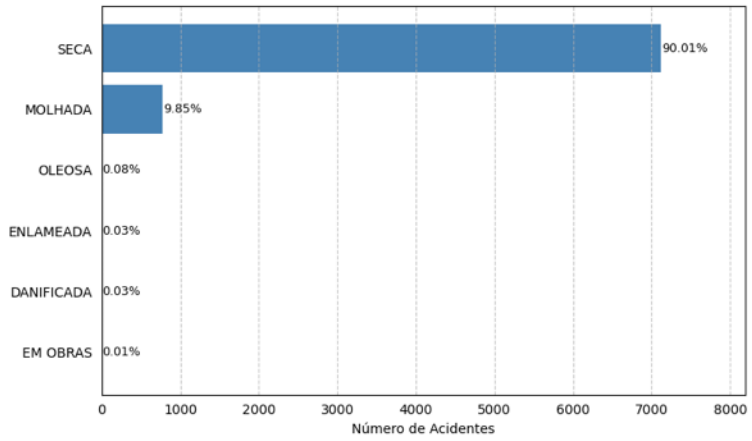


Figura 11 - Sinistros por condição da via.



Caracterização dos Envolvidos nos Sinistros

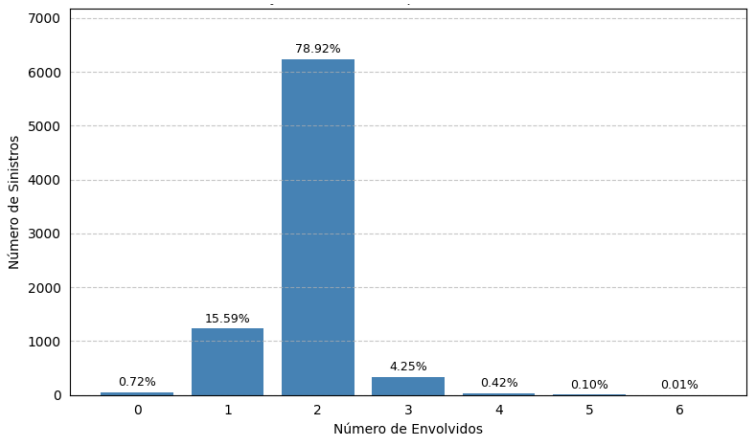
Com base na distribuição do número de envolvidos por sinistro, observou-se que a maioria dos registros refere a eventos com dois envolvidos, totalizando 6.253 ocorrências (78,92%). Essa predominância é compatível com colisões envolvendo dois veículos, o que é típico em zonas urbanas.

Em segundo lugar, estão os sinistros com apenas um envolvido, que somam 1.235 registros (15,59%). Esses casos podem estar relacionados, por exemplo, a situações de perda de controle do veículo, choques em objetos fixos ou quedas de motociclistas sem

outro veículo envolvido. Já os registros com três envolvidos representam 336 ocorrências (4,25%). Casos com quatro ou mais envolvidos são raros, com apenas 33 registros (0,6%).

Um ponto que merece atenção é a presença de 58 registros com zero envolvidos declarados (0,72%) do total. Essa situação provavelmente refere-se a sinistros em que os condutores evadiram do local antes da chegada da autoridade responsável pelo registro, impossibilitando a identificação dos envolvidos no momento da coleta dos dados. A Figura12 apresenta graficamente estes dados.

Figura 12 - Distribuição dos Sinistros por Número de Envolvidos.



A Tabela 3 apresenta a análise cruzada entre o número de envolvidos e a natureza do sinistro, revelando que a maioria das ocorrências classificadas como colisão envolveu dois participantes, com 2.896 registros, o que representa 83,89% dos casos dessa categoria. Situação semelhante é observada nos abalroamento, com 2.604 ocorrências (84,55%) com dois envolvidos.

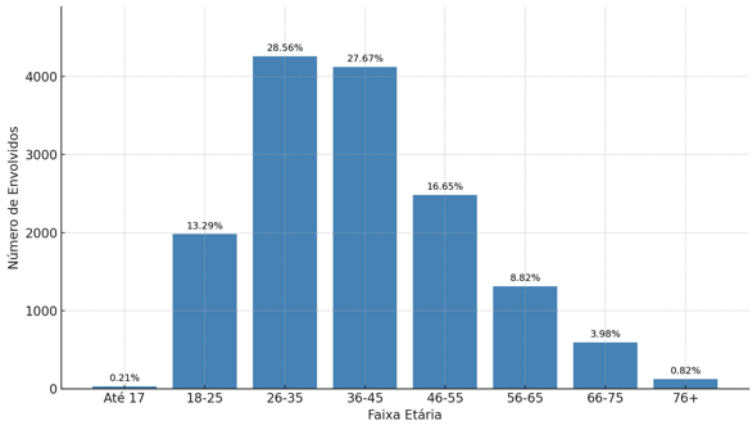
Nos choques em objeto fixo, há maior dispersão: 53,21% ocorrem com dois envolvidos e 38,06% com apenas um, sugerindo perda de controle do veículo. Nos atropelamentos, 76,92% dos registros envolvem dois participantes, e 19,23% apenas um. Os tombamentos e capotamentos ocorrem, majoritariamente, com um único envolvido (64,29% e 50%, respectivamente), compatível com eventos em que apenas um veículo está em movimento. Por fim, os atropelamentos de animal aparecem na maioria com apenas um envolvido (75%).

Tabela 3 – Distribuição dos sinistros por natureza e nº de envolvidos.

Natureza / Envolvidos	0	1	2	3	4	5	6
Colisão	8 (0.23%)	320 (9.27%)	2896 (83.89%)	201 (5.82%)	23 (0.67%)	4 (0.12%)	0 (0%)
Abalroamento	0 (0.0%)	402 (13.05%)	2604 (84.55%)	69 (2.24%)	3 (0.1%)	2 (0.06%)	0 (0%)
Choque em objeto fixo	43 (3.83%)	427 (38.06%)	597 (53.21%)	47 (4.19%)	5 (0.45%)	2 (0.18%)	1 (0.09%)
Outro	4 (2.26%)	57 (32.2%)	99 (55.93%)	15 (8.47%)	2 (1.13%)	0 (0%)	0 (0%)
Atropelamento	0 (0.0%)	10 (19.23%)	40 (76.92%)	2 (3.85%)	0 (0%)	0 (0%)	0 (0%)
Tombamento	1 (7.14%)	9 (64.29%)	4 (28.57%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Capotamento	1 (10.0%)	5 (50.0%)	2 (20.0%)	2 (20.0%)	0 (0%)	0 (0%)	0 (0%)
Atropelamento animal	0 (0.0%)	3 (75.0%)	1 (25.0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

A distribuição por faixa etária dos envolvidos em sinistros (Figura 13) mostra concentração nas idades entre 18 e 35 anos, que correspondem à maioria dos registros. As faixas de 18 a 25 e de 26 a 35 anos lideram em número absoluto, refletindo a maior exposição desse grupo à condução de veículos e a comportamentos de maior risco no trânsito. As faixas entre 36 e 55 anos também apresentam participação relevante, enquanto os registros diminuem progressivamente a partir dos 56 anos. Casos envolvendo menores de 18 anos são pouco frequentes, possivelmente por não estarem habilitados ou formalmente identificados como condutores.

Figura 13 - Distribuição por Faixa Etária dos Envolvidos nos Sinistros.



A distribuição por sexo dos envolvidos nos sinistros indica que 10.792 registros correspondem ao sexo masculino, o que representa 72,0% do total, enquanto 4.197

registros são do sexo feminino, correspondendo a 28,0%. A soma total de registros (14.989) ultrapassa o número de sinistros registrados, uma vez que em grande parte dos eventos há mais de um envolvido.

O cruzamento entre sexo e natureza do sinistro revela que, independentemente do tipo de ocorrência, os homens representam a maioria dos envolvidos em todos os tipos de sinistros registrados.

Nos abalroamentos e colisões, que concentram o maior volume de registros, as mulheres correspondem a 28,39% e 29,01% dos casos, respectivamente, enquanto os homens estão presentes em mais de 70% dessas ocorrências.

Em sinistros do tipo choque em objeto fixo, a participação feminina é ainda menor (23,7%), indicando possível associação com perda de controle veicular em contextos em que a condução é predominantemente masculina.

Nos atropelamentos, as mulheres estão envolvidas em 26,8% dos casos, enquanto nos tombamentos e atropelamentos de animal, não há registro feminino, com 100% dos envolvidos sendo homens.

Análise de Alcoolemia

A realização do teste de alcoolemia é um procedimento fundamental para a caracterização adequada de sinistros de trânsito, permitindo identificar a presença de substâncias psicoativas que podem comprometer a capacidade de condução. O consumo de álcool é amplamente reconhecido como fator de risco associado ao aumento da gravidade e da ocorrência de sinistros, afetando reflexos, tempo de reação e julgamento do condutor. A testagem sistemática, além de fornecer subsídios técnicos para a responsabilização legal, contribui para o monitoramento de padrões de comportamento no trânsito e para a formulação de políticas públicas de prevenção, especialmente no que se refere à fiscalização e à educação para a segurança viária.

Os dados indicam que entre os 7.911 registros de sinistros, 527 condutores (6,66%) realizaram o teste de alcoolemia no local. Em 7.375 dos casos (93,22%), não teve o teste realizado, enquanto 9 condutores (0,11%) recusaram a fazê-lo.

Analisando a distribuição dos testes de alcoolemia realizados por dia da semana, observa-se que os testes são mais frequentemente aplicados nos finais de semana, especialmente, com destaque para o ano de 2022, que registrou 51 testes realizados aos sábados e 35 aos domingos, superando os demais anos nesse recorte (Figura 14).

Em relação aos turnos do dia, o período noturno concentra a maior parte dos testes em todos os anos, com destaque para 2022, quando foram registrados 108 testes, seguido de 74 em 2023 e 59 em 2024. O turno da tarde aparece como o segundo período com mais testes realizados, mantendo certa estabilidade ao longo dos anos, entre 32 e 35 registros entre 2022 e 2024. Na madrugada, o número de testes manteve-se relativamente estável entre 2021 e 2023, com leve queda em 2024 (de 17 para 11). Apesar da menor quantidade, esse turno ainda é relevante para fins de fiscalização, considerando o potencial risco associado à condução sob influência de álcool nesse horário (Figura 15).

Figura 14 - Testes de Alcoolemia Realizados no Local por Dia da Semana e Ano.

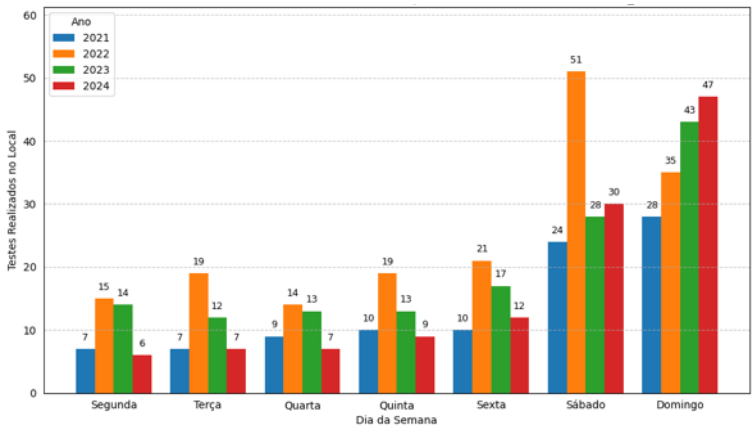
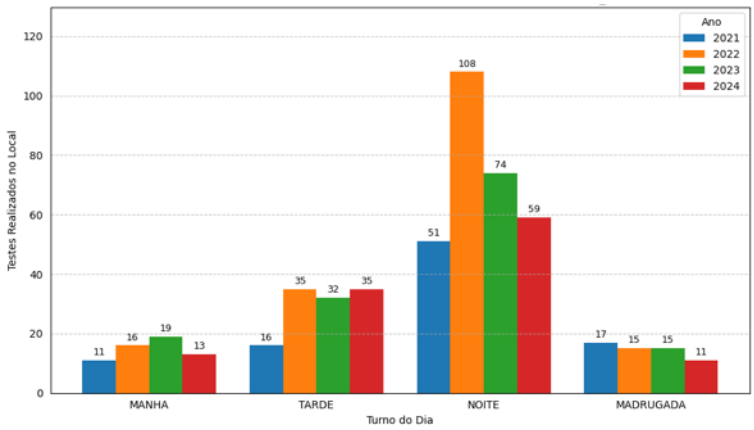


Figura 15 - Testes de Alcoolemia Realizados no Local por Turno e Ano.



Segundo determinação do Conselho Nacional de Trânsito (Contran), conforme a Resolução nº 432/2013, são estabelecidos os intervalos de alcoolemia e as respectivas implicações legais. Os valores até 0,04 mg/L indicam liberação do condutor; os valores compreendidos entre 0,05 mg/L e 0,33 mg/L caracterizam infração gravíssima; e valores

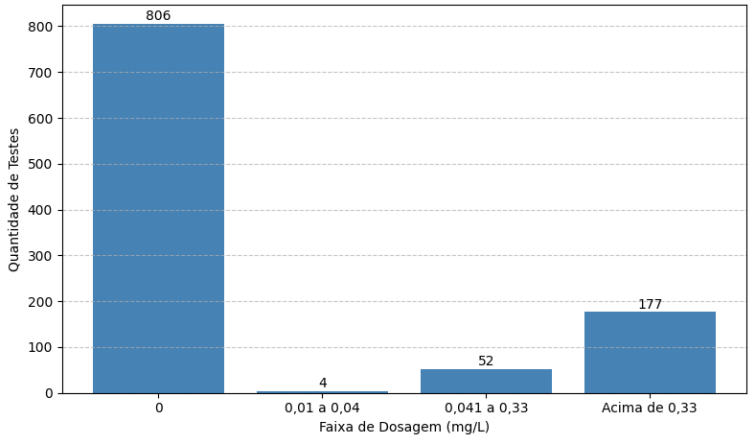
iguais ou superiores a 0,34 mg/L configuram crime de trânsito, com penalidades mais severas.

A Figura 16 apresenta a distribuição dos testes de alcoolemia por faixa de dosagem, o que evidencia que a maior parte dos condutores testados apresentaram resultado igual a zero. Foram 806 testes (cerca de 77%) com 0 mg/L, o que indica ausência de álcool no momento da abordagem. Os valores referem-se a todos os testes realizados, levando em consideração que há mais de um envolvido em alguns registros de sinistro.

A segunda maior faixa é a de dosagens acima de 0,33 mg/L, com 177 registros, o que representa um percentual relevante entre os testes positivos. Essa concentração sugere que, entre os condutores que apresentaram presença de álcool, muitos estavam acima do limite legal de tolerância, configurando infração gravíssima segundo o Código de Trânsito Brasileiro.

A faixa intermediária de 0,041 a 0,33 mg/L reuniu 52 casos, e a de 0,01 a 0,04 mg/L, apenas 4 registros, indicando que resultados próximos ao limite de detecção ou em níveis baixos foram menos frequentes.

Figura 16 - Distribuição dos Testes de Alcoolemia por Faixa de Dosagem.



3.4. Conclusão

Este estudo examinou 7.911 sinistros de trânsito, selecionados de um universo de 7.926 registros eletrônicos gerados pela Agência Municipal de Mobilidade e Trânsito de Rio Verde entre 2021 e 2024. A coleta digital reuniu informações padronizadas e georreferenciadas, reduzindo o esforço de tratamento dos dados, minimizou erros de

preenchimento e elevou a qualidade global da base. No âmbito metodológico, o emprego de coleta eletrônica demonstrou eficaz em reduzir substancialmente erros de preenchimento e simplificar o pré-processamento, gerando uma base de alta qualidade apta a análises reprodutíveis. Em consequência, foi possível realizar análises temporais e espaciais com maior acurácia, além de demonstrar que todo o fluxo de atualização pode ser automatizado sem intervenção manual, liberando os dados quase em tempo real.

Os sinistros concentram-se nos eixos viários dos bairros Setor Central, Jardim Goiás e Bairro Popular; apresentam distribuição anual relativamente estável e atingem pico nas tardes de dias úteis. Colisões entre veículos respondem por cerca de 44%, seguidas de abalroamentos laterais e choques em objeto fixo. O perfil predominante dos envolvidos corresponde a homens entre 18 e 35 anos.

Os dados mostraram distribuição anual estável, variando de 1.834 ocorrências em 2021 a 2.094 em 2024, com pico vespertino em dias úteis. No âmbito espacial, verificou-se concentração de eventos nos bairros Setor Central, Jardim Goiás e Bairro Popular, resultado que se tornou evidente graças às coordenadas armazenadas em cada registro. As tipologias predominantes foram colisões entre veículos, que representaram aproximadamente 44% do total, seguidas de abalroamentos laterais e choques em objeto fixo; juntas, essas três categorias responderam por mais de 95% dos sinistros. Quanto ao perfil de vítimas e condutores, a maioria era composta por homens com idade entre 18 e 35 anos, e apenas 6,7% dos condutores submetidos ao teste de alcoolemia apresentaram resultado positivo.

Esses achados confirmam que a coleta eletrônica garante dados de alta qualidade e reforçam, em termos teóricos, a evidência de que o risco viário se concentra em corredores centrais de cidades médias brasileiras, reduzindo a necessidade de tratamento manual e permite análises mais acuradas.

Recomenda-se integrar esses registros à base da Polícia Militar, que contém dados detalhados sobre vítimas, a fim de ampliar o conjunto de variáveis clínicas e fortalecer futuras análises preditivas. A combinação de dados de alta qualidade, localização precisa e integração interinstitucional cria condições favoráveis para políticas de segurança viária fundamentadas em evidências quantitativas. Sugere-se, ainda, integrar dados hospitalares, desenvolver modelos espaço-temporais preditivos e avaliar intervenções de engenharia, fiscalização ou educação por meio de desenhos antes-e-depois ou controle sintético. A padronização periódica das tipologias, para reduzir a categoria ‘Outro’, complementa a agenda de melhoria contínua.

3.5. Referências Bibliográficas

DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TRANSPORTES (DNIT). Relatório de **sinistros de trânsito.** Disponível em: https://www.gov.br/dnit/pt-br/assuntos/infraestrutura-rodoviaria/sinistros-de-transito. Acesso em: 10 fev. 2025.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). NBR 10697: **Segurança no trânsito – Terminologia**. Rio de Janeiro, 2020.

BRASIL. Ministério dos Transportes. **Plano Nacional de Redução de Mortes e Lesões no Trânsito (PNATRANS)**. Disponível em: https://www.gov.br/transportes/pt-br/assuntos/transito/pnatrans. Acesso em: 10 fev. 2025.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS). **Plano Global - Década de Ação pela segurança no trânsito 2021-2030**. Disponível em: https://cdn.who.int/media/docs/default-source/documents/health-topics/road-traffic-injuries/global-plan-for-the-doa-of-road-safety-2021-2030-pt.pdf?sfvrsn=65cf34c8_35&download=true. Acesso em: 11 fev. 2025.

BACCHIERI, G.; BARROS, A. J. D. **Acidentes de trânsito no Brasil de 1998 a 2010: muitas mudanças e poucos resultados**. Revista de Saúde Pública, v. 45, n. 5, p. 949-963, 2011.

CONFEDERAÇÃO NACIONAL DE MUNICÍPIOS (CNM). **Mapeamento das mortes por acidentes de trânsito**. Brasília: CNM, 2012. Disponível em: https://cnm.org.br/biblioteca/download/411. Acesso em: 10 fev. 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Década de Ação pela Segurança no Trânsito 2021-2030**. Disponível em: https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/decade-of-action-for-road-safety-2021-2030. Acesso em: 12 fev. 2025.

HAUER, E.; HAKKERT, A. S. **Extent and Some Implications of Incomplete Accident Reporting**. Transportation Research Record, n. 1185, p. 1-10, 1988.

MELLO-JORGE, M. H. P. **Situação atual das estatísticas oficiais relativas à mortalidade por causas externas**. Revista de Saúde Pública, v. 24, n. 3, p. 217-223, 1990.

ALSOP, J.; LANGLEY, J. **Under-reporting of motor vehicle traffic crash victims in New Zealand**. Accident Analysis & Prevention, v. 33, n. 3, p. 353-359, 2001.

SALIFU, M.; ACKAAH, W. **Under-reporting of road traffic crash data in Ghana.** International Journal of Injury Control and Safety Promotion, v. 19, n. 4, p. 331-339, 2012.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. **Estatística: teoria e aplicações usando Microsoft Excel em português**. 5. ed. Rio de Janeiro: LTC, 2016.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2005.

MAHANTI, R. **Data Quality: Dimensions, Measurement, Strategy, Management, and Governance**. Milwaukee: ASQ Quality Press, 2019.

LONCZAK, H. S.; NEIGHBORS, C.; DONOVAN, D. M. **Predicting risky and angry driving as a function of gender**. Accident Analysis & Prevention, v. 39, n. 3, p. 536-545, 2007.

4. CAPÍTULO II

SINISTROS DE TRÂNSITO EM RIO VERDE (GO): ABORDAGEM ESPACIAL COM INDICADORES DE AUTOCORRELAÇÃO E AGRUPAMENTO

RESUMO

O estudo avaliou 6.857 sinistros de trânsito sem vítimas ocorridos em Rio Verde (GO) entre 2021 e 2024. Empregaram-se estimativa de densidade por *kernel* (KDE) para representar padrões contínuos, índice de Moran Global e Indicadores Locais de Associação Espacial (LISA) para medir autocorrelação e o algoritmo DBSCAN para análise linear do principal corredor viário da cidade. A KDE apontou aglomeração persistente de ocorrências na Avenida Presidente Vargas e nas rodovias BR-452/GO-174; gradientes decrescentes foram observados nas zonas periféricas. O Moran Global indicou dependência espacial positiva significativa (I = 0,5897; p = 0,001), enquanto o LISA evidenciou *clusters* Alto-Alto coincidentes com a malha arterial central. No recorte da Avenida Presidente Vargas, o DBSCAN isolou 353 sinistros distribuídos ao longo de 7,23 km, destacando trechos críticos. Os resultados reforçam a relação entre hierarquia viária e risco e sustentam recomendações de intervenções de engenharia, fiscalização direcionada ao modal leve e adoção de painéis geoespaciais para monitoramento contínuo, alinhadas às metas do PNATRANS e da Década 2021-2030.

Palavras-chave: sinistros de trânsito; análise espacial; Moran I; KDE; Rio Verde-GO.

ABSTRACT

This study analyzed 6,857 non-injury traffic crashes recorded in Rio Verde (GO), Brazil, between 2021 and 2024. Kernel Density Estimation (KDE) was applied to identify continuous spatial patterns, while Global Moran's I and Local Indicators of Spatial Association (LISA) were used to measure spatial autocorrelation. The DBSCAN algorithm was employed to analyze the linear distribution of crashes along the city's main traffic corridor. KDE revealed a persistent concentration of crashes along Presidente Vargas Avenue and highways BR-452/GO-174, with decreasing gradients toward peripheral areas. Global Moran's I indicated significant positive spatial dependence (I = 0.5897; p = 0.001), and LISA identified High-High clusters aligned with the central arterial road network. In the specific analysis of Presidente Vargas Avenue, DBSCAN isolated 353 crashes distributed along 7.23 km, highlighting the critical segments. The findings reinforce the relationship between road hierarchy and crash risk and support recommendations for engineering interventions, targeted enforcement on light-duty vehicles, and the adoption of geospatial dashboards for continuous monitoring—aligned with the goals of PNATRANS and the UN Decade of Action for Road Safety 2021–2030.

Keywords: traffic crashes; spatial analysis; Moran's I; KDE; Rio Verde (Brazil).

4.1. Introdução

A segurança viária permanece um problema de saúde pública global, responsável por cerca de 1,19 milhão de mortes anuais, sobretudo em países de renda média e baixa (OMS, 2023). Reconhecendo a gravidade da situação, a Resolução 74/299 da Assembleia Geral das Nações Unidas instituiu a Década de Ação pela Segurança no Trânsito 2021-2030, estabelece a meta de reduzir em pelo menos 50% as mortes e lesões no trânsito nesse intervalo (ONU, 2020).

No cenário internacional, o Brasil é o terceiro país no mundo com mais mortes no trânsito, atrás apenas de Índia e a China.

Em 2024, o Registro Nacional de Sinistros e Estatísticas de Trânsito (RENAEST) contabilizou 1.140.114 sinistros de trânsito no Brasil, envolvendo 1.689.779 veículos e resultando em 1.414.873 pessoas feridas ou ilesas e 21.525 óbitos. Esses números correspondem a taxa de 10,04 mortes por 100 mil habitantes e a 1,74 mortes por 10 mil veículos, com óbitos representando 1,89% do total de sinistros registrados no período. Ainda que a análise deste estudo se concentre nos sinistros sem vítimas, esses dados fornecem um panorama geral da magnitude do fenômeno no país, evidenciando o impacto tanto em termos absolutos quanto proporcionais.

Esses indicadores evidenciam a magnitude do problema no país e reforçam a urgência de estratégias fundamentadas em precisão metodológica e terminológica para orientar políticas públicas eficazes.

Diante desse cenário, o Plano Nacional de Redução de Mortes e Lesões no Trânsito (PNATRANS) alinha-se a essa orientação internacional e embasa políticas públicas voltadas à mobilidade segura. O alinhamento dessas estratégias globais e nacionais depende também de precisão conceitual e uniformidade terminológica, fatores essenciais para a análise e comparação de dados de sinistros. Em consonância com a revisão da ABNT NBR 10697:2018, que substituiu o termo “acidente de trânsito” por “sinistro de trânsito” para reforçar a compreensão de que a maioria desses eventos é evitável e não fruto de mero acaso, este estudo adota essa terminologia, destacando a perspectiva de que tais ocorrências são preveníveis e passíveis de mitigação por meio de políticas públicas eficazes.

Embora capitais brasileiras já contem com diagnósticos espaciais consolidados (MELO; MENDONÇA, 2021; PAIXÃO et al., 2015; SILVA; PEREIRA; ALVES, 2021),

municípios médios do interior, cujas malhas viária e urbana convivem com tráfego agroindustrial intenso, carecem de investigações equivalentes. Rio Verde (GO), polo regional do Sudoeste Goiano, apresenta frota total de 189.862 veículos e população estimada em 237.092 habitantes (IBGE, 2022). Dados do Relatório Estatístico de Sinistros de Trânsito (RENAEST/SENATRAN) indicam, para o período 2021-2024, 15.824 sinistros registrados, que envolveram 22.523 veículos, resultaram em 23.077 pessoas feridas ou ilesas e 102 óbitos, resultando em coeficiente de 43,02 óbitos/100.000 habitantes e 5,37 óbitos/10 000 veículos (SENATRAN, 2024).

Diante desse contexto, buscou-se analisar a distribuição espacial e a evolução temporal dos sinistros de trânsito em Rio Verde (GO) entre 2021 e 2024, empregando KDE, Moran I/LISA e DBSCAN. Especificamente, buscou-se: (i) caracterizar a variação espaço-temporal dos sinistros em escala municipal e intraurbana; (ii) identificar *hotspots* e corredores críticos, com destaque para o eixo da Avenida Presidente Vargas; e (iii) discutir implicações para políticas públicas locais de segurança viária, em consonância com as metas do PNATRANS e da Década 2021-2030.

Ferramentas de análise espacial têm demonstrado elevado potencial para revelar padrões de distribuição de sinistros, identificar *hotspots* e respaldar intervenções baseadas em evidências. Estudos como Plug, Xia e Calfield (2011), Shafabakhsh, Famili e Bahadori (2017) e Munasinghe (2023) aplicaram, respectivamente, Estimativa de Densidade por *Kernel* (KDE), estatísticas globais e locais de autocorrelação (Moran I e LISA) e o algoritmo DBSCAN para delimitação de áreas críticas, fornecendo referenciais metodológicos consistentes.

O presente trabalho contribui para suprir lacuna na literatura sobre cidades intermediárias brasileiras e fornece subsídios técnicos ao planejamento urbano-viário de Rio Verde, favorecendo a priorização de investimentos em infraestrutura segura, fiscalização e educação para o trânsito.

4.2. Material e Método

4.2.1. Área de estudo

Rio Verde localiza-se no Sudoeste Goiano (17°47'53"S; 50°55'15"W), apresentando área de 8.386 km² e população estimada em 237.092 habitantes (IBGE, 2022). A malha urbana é atravessada por rodovias de escoamento agroindustrial (BR-060, BR-452 e GO-174), condição que eleva o volume de tráfego pesado e o potencial de conflitos viários. O relevo suavemente ondulado, aliado à rápida expansão urbana, gera vias com geometrias diversas e variação de controle de acesso, elementos relevantes para os sinistros de trânsito.

4.2.2. Coleta e pré-processamento de dados

A base de dados foi constituída a partir de planilha de sinistros sem vítimas, encaminhada pela Agência Municipal de Mobilidade e Trânsito (AMT), de Rio Verde, contendo 7.926 registros compreendidos entre 01 de janeiro de 2021 e 31 de dezembro de 2024. Cada registro apresenta número do boletim, agente responsável, solicitante, endereço do fato, coordenada geográfica (SIRGAS 2000) e link para o formulário eletrônico do boletim. O formulário, por sua vez, armazena data, hora, dados do sinistro, condições meteorológicas, estado da sinalização e do pavimento, imagens, relatos e classificação de danos veiculares. Os registros são preenchidos por agentes da AMT por meio de formulários eletrônicos, armazenados em banco de dados e disponibilizados internamente no formato HTML. Ressalta-se que esse recorte decorre da disponibilidade institucional: os registros de sinistros com vítimas são de responsabilidade de outros órgãos e não foram acessíveis para esta pesquisa.

Pré-processamento dos dados

Conceitualmente, o pré-processamento de dados abrange rotinas de limpeza, padronização e integração de informações oriundas de múltiplas fontes, assegurando coerência semântica e espacial antes de qualquer modelagem ou inferência. Estudos de

referência evidenciam que falhas nessa etapa comprometem a fidedignidade dos indicadores de risco e a localização de *hotspots*: Montella (2010) demonstrou que a padronização de códigos viários aprimora a identificação de segmentos críticos, e Xie e Yan (2013) reportaram maior sensibilidade na detecção de *clusters,* após a harmonização entre bases policiais e operacionais.

Para o pré-processamento, análise estatística e representação espacial dos dados neste estudo, empregou-se a linguagem Python, comumente utilizada em estudos de mobilidade urbana e segurança viária (Çalişkan; Anbaroğlu, 2023). As principais bibliotecas utilizadas foram: *requests* e *BeautifulSoup* para *web scraping* (automação de acesso e extração de conteúdo); pandas e *numpy* para manipulação de dados; *matplotlib* e *seaborn* para visualizações gráficas; *math* para operações matemáticas básicas; e *geopandas, folium* e *contextily* para manipulação e visualização de dados geográficos.

À luz desse corpo de evidências e considerando os requisitos técnicos empregados, estruturou-se neste estudo um *pipeline* de pré-processamento composto por três etapas, descritas a seguir.

- **Integração dos dados:** Para extrair e estruturar informações detalhadas que não estavam diretamente disponíveis na base principal, mas acessíveis por meio de formulários eletrônicos via link, foi desenvolvido um processo automatizado utilizando as bibliotecas pandas, *requests* e *BeautifulSoup*. Este processo constituiu na raspagem, acessando cada URL do boletim, extraindo o conteúdo em texto simples, com armazenamento em arquivos txt, identificados por um código único. Após, os arquivos foram lidos e convertidos em pares chave-valor, que geraram dois *datasets*: geral dos sinistros e dados dos envolvidos. Por fim, foi realizada a integração para formato largo e a função *join* entre as tabelas, eliminando duplicidades e gerando um *dataset* único.
- **Transformação:** O procedimento de transformação consistiu na adaptação e reestruturação das variáveis para adequação aos métodos de análise estatística e espacial. Destacam-se a formatação conjunta das variáveis de data e hora no padrão *datetime*, a extração dos valores de latitude e longitude a partir da coluna de coordenadas geográficas (convertidos para o tipo numérico apropriado) e a categorização de variáveis textuais recorrentes, como turno, bairro e natureza do sinistro, para o tipo categórico.

- **Limpeza:** Foram excluídos os registros que não continham formulário eletrônico associado, que apresentavam ausência de coordenadas geográficas ou cujas coordenadas estavam inconsistentes, (casos em que o ponto registrado correspondia à sede da AMT, apesar de o endereço informado remeter ao local real do fato).

Após as etapas de integração, tratamento e limpeza, os dados resultantes foram submetidos a um processo de validação por verificação manual. Uma amostra de registros foi inspecionada e comparada com os boletins eletrônicos originais, a fim de confirmar a consistência das informações extraídas e armazenadas. Essa etapa visou mitigar o risco de distorções introduzidas pelo processamento automático e garantir a fidedignidade dos dados empregados nas análises subsequentes.

Essa etapa do pré-processamento resultou em redução progressiva da base original, que passou de 7.926 para 6.857 registros válidos, conforme detalhado na Tabela 4.

Tabela 4 - Etapas de filtragem da base de dados.

Etapas de filtragem da base de dados	Registros removidos	Total de registros remanescentes
Base de dados inicial	-	7.926
Remoção de registros sem formulários	15	7.911
Remoção de registros sem coordenadas	163	7.748
Remoção de registros com coordenadas inconsistentes	891	6.857

4.2.3. Técnicas de análise espaciotemporal

Estimativa de Densidade por *Kernel*

A Estimativa de Densidade por *Kernel* é reconhecida como uma das técnicas mais eficazes para representar padrões espaciais contínuos a partir de eventos pontuais, como sinistros de trânsito (Chainey; Ratcliffe, 2013; O'Sullivan; Unwin, 2003). O método consiste em sobrepor uma função núcleo simétrica a cada ocorrência e, em seguida, somar as contribuições individuais para gerar uma superfície de densidade, permitindo identificar *hotspots* em que a concentração de eventos é estatisticamente mais elevada (FOTHERINGHAM et al., 2000).

Do ponto de vista matemático, a KDE estima a densidade $\hat{f}(x)$ em cada localização x conforme:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\left||x - X_i|\right|}{h}\right)$$

em que n é o número de observações, h o parâmetro de suavização (bandwidth), d a dimensão espacial, K a função núcleo e $||x - X\_i||$ a distância euclidiana entre a posição x e o ponto i. Neste estudo, empregou-se o núcleo *quartic* pelo bom desempenho computacional e capacidade de suavização (SILVERMAN, 1986).

A seleção de h revela-se crítica: larguras de banda elevadas produzem superfícies excessivamente suavizadas que ocultam *clusters* locais, ao passo que valores muito pequenos geram ruído visual e estatístico. Métodos iterativos e de validação cruzada são indicados para calibrar h (Xie; Yan, 2008; SHARIAT-MOHAYMANY; KHAKPOOR; KESHTKAR, 2013). Após testes entre 100m e 1.000m, adotou-se 800m por oferecer equilíbrio entre detalhamento e robustez para a malha urbana de Rio Verde.

A seleção de h é um aspecto determinante na suavização das superfícies estimadas. Larguras de banda mais elevadas produzem mapas excessivamente alisados, ocultando concentrações locais, enquanto valores muito baixos geram ruído visual e estatístico. Neste estudo, a estimação foi conduzida em coordenadas geográficas (WGS84, EPSG:4326), adotando-se um parâmetro de suavização relativo bw_method = 0,15, que corresponde aproximadamente a 0,15° em latitude (cerca de 16–17 km). A escolha foi feita a partir de testes comparativos, buscando equilíbrio entre legibilidade e robustez das estimativas no contexto urbano analisado.

Índice de Moran Global e Local (LISA)

O índice de Moran é uma estatística empregada na análise de autocorrelação espacial, permitindo identificar se padrões espaciais de uma variável são aleatórios, agrupados ou dispersos (ANSELIN, 1995; GETIS, 2007). Na análise de sinistros de trânsito, sua aplicação permite verificar se as ocorrências apresentam dependência espacial, isto é, se há agrupamentos de áreas com taxas elevadas ou baixas de eventos próximos entre si (Xie; Yan, 2008; Shariatmohaymany *et al*., 2013).

A versão global do índice de Moran é expressa pela fórmula:

$$I = \frac{n}{W} \cdot \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

em que n representa o número de áreas, x_i o valor observado na área i, $\bar{x}$ a média global, w_{ij} os pesos espaciais entre as áreas i e j, e W a soma total dos pesos. Valores de $I > 0$ indicam autocorrelação positiva (agrupamento de valores semelhantes), $I < 0$ indicam autocorrelação negativa (valores diferentes próximos) e $I \approx 0$ indicam aleatoriedade espacial (Anselin, 1995).

Como complemento à estatística global, os Indicadores Locais de Associação Espacial (LISA) permitem decompor o índice de Moran em componentes locais, possibilitando identificar *clusters* específicos, como agrupamentos de altas taxas (*High-High*), baixas taxas (*Low-Low*) ou *outliers* espaciais (*High-Low e Low-High*). A estatística LISA é expressa por:

$$I_i = \frac{(x_i - \bar{x})}{m_2} \sum_j w_{ij}(x_j - \bar{x})$$

em que I_i representa o índice local para a área i, m_2 é a variância da variável e os demais termos seguem a definição anterior. Esses indicadores são representados graficamente por mapas de autocorrelação espacial, cuja significância é comumente testada por permutações aleatórias (GETIS, 2007).

***Density-Based Spatial Clustering of Applications with Noise* (DBSCAN)**

O algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) é um método de agrupamento baseado em densidade proposto por ESTER *et al*. (1996), utilizado para identificação de *clusters* espaciais em contextos urbanos com dados ruidosos. A lógica baseia-se na formação de regiões densamente conectadas, com a capacidade de isolar pontos dispersos considerados ruídos. A definição dos agrupamentos depende de dois parâmetros principais: ε (*epsilon*), que representa o raio de vizinhança de busca, e *min_samples*, que é o número mínimo de pontos exigido nesse raio para que um ponto seja considerado núcleo de um *cluster*.

Formalmente, um ponto p é classificado como núcleo se existir um conjunto de pelo menos minPts pontos $q_1, q_2, \ldots, q_{minPts}$ tal que a distância entre p e cada q_i satisfaça

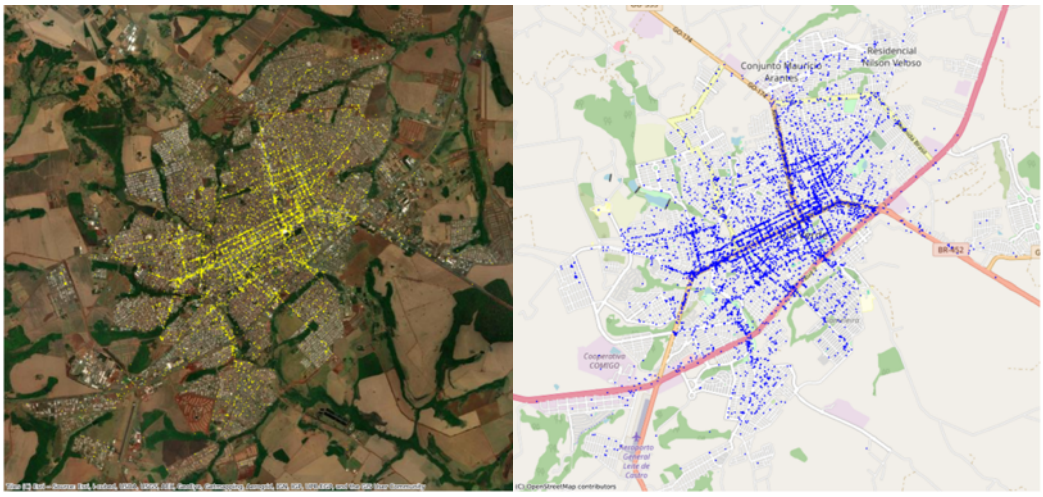
$\text{dist}(p, q_i) \leq \varepsilon$. O agrupamento forma-se pela transitividade das conexões entre pontos núcleo e os vizinhos densamente conectados. Os pontos que não pertencem a nenhum agrupamento denso são classificados como *outliers* espaciais. Essa abordagem torna o DBSCAN robusto frente a variações de forma e densidade dos agrupamentos, sendo especialmente indicado para análise de dados geográficos com distribuição irregular (PRASANNAKUMAR *et al*., 2018).

A principal aplicação em análises espaciais está na detecção de padrões agrupados de eventos, utilizado neste estudo permitindo isolar regiões com concentração significativa de ocorrências, sem impor geometrias pré-definidas aos *clusters*. A sensibilidade dos resultados à escolha dos parâmetros torna recomendável o uso de procedimentos exploratórios para calibração de ε e *minPts*, considerando a escala espacial do fenômeno analisado.

4.3. Resultados e Discussão

A Figura 17 apresenta a distribuição espacial dos 6.857 sinistros sem vítimas registrados pela AMT entre 2021 e 2024, exibida sobre dois fundos cartográficos distintos: um mosaico de imagens de satélite e a camada vetorial do OpenStreetMap. Em ambas as visualizações, os pontos concentram-se ao longo da malha arterial, com ênfase nos corredores BR-452/GO-174, na Avenida Presidente Vargas e nos eixos que conectam o centro urbano aos distritos industriais. A densidade de ocorrências decresce também à medida que se afastam os acessos à BR-060, confirmando a correlação entre hierarquia viária e risco de sinistros descrita na literatura especializada (PLUG *et al*., 2011; XIE; YAN, 2013).

Figura 17 - Satélite e camada vetorial - 2021 a 2024.

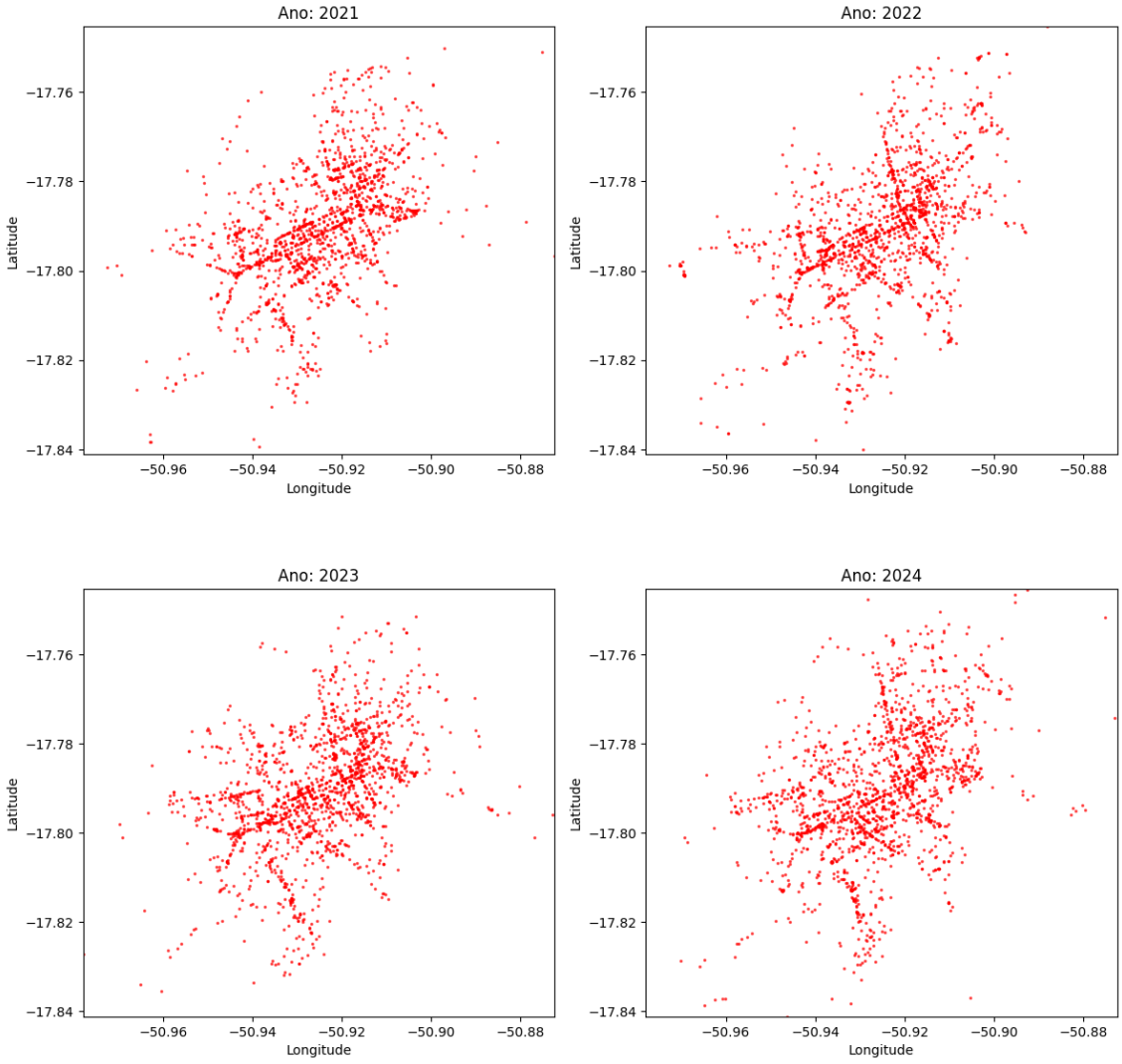


A Figura 18 sintetiza a distribuição espaço-temporal dos sinistros ao longo dos quatro anos. Em todos os quadros persiste um núcleo de alta concentração na confluência dos eixos Presidente Vargas – BR-452/GO-174, indicando estabilidade do padrão espacial central ao longo do quadriênio.

Três padrões destacam-se:

- **Persistência do eixo central:** Em todos os quadros observa-se adensamento contínuo ao longo do corredor Avenida Presidente Vargas – BR-452/GO-174, corroborando a associação entre hierarquia viária e risco descrita por PLUG *et al*. (2011) e XIE E YAN (2013).
- **Estabilidade periférica:** A baixa dispersão nos quadrantes oeste e sul indica que intervenções locais (alargamento de faixas, redutores) não alteraram substancialmente o macro-padrão espacial.
- **Oscilações discretas no quadrante nordeste:** Pontos adicionais surgem em 2022, retraem-se em 2023 e 2024, sugerindo efeito pontual de obras viárias ou variação de fluxo; a inexistência de “drift” sistemático do centroide anual reforça a hipótese de concentração estrutural.

Figura 18 – Distribuição temporal 2021-2024.



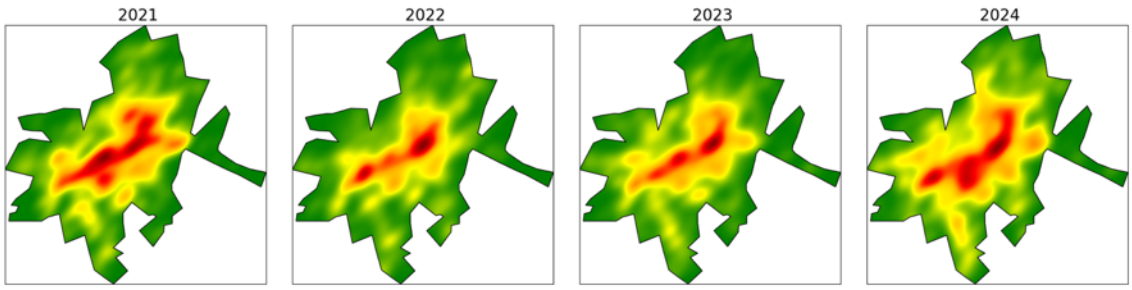
Estimativa de Densidade de *Kernel* (KDE)

Para estimar a variação espaço-temporal dos mapas desta seção, aplicou-se a estimativa de densidade por *kernel* (KDE) com os seguintes parâmetros: sistema de referência geográfica WGS 84 (EPSG 4326); malha de interpolação regular de 500 × 500 nós sobre o envelope municipal; núcleo gaussiano bivariado implementado em *gaussian_kde*; largura de banda fixa igual a 0,15 de *h* de Scott, resultando em raio efetivo aproximado de 250–400m (Silverman, 1986); máscara espacial mediante o polígono oficial do perímetro urbano; e paleta contínua “INPE” (verde-amarelo-laranja-vermelho) para realçar gradientes de densidade.

A inspeção conjunta dos quatro painéis da Figura 19 confirma a existência de um *hotspot* persistente, alinhado ao corredor Avenida Presidente Vargas, cujo núcleo vermelho se mantém em todas as superfícies. No total, foram registrados 1.642 sinistros

em 2021 (23,9%), 1.681 em 2022 (24,5%), 1.741 em 2023 (25,4%) e 1.793 em 2024 (26,1%), totalizando as 6.857 ocorrências no quadriênio. Em 2021, observam-se focos secundários no sudoeste, próximos ao entroncamento com a GO-174, e uma mancha menos intensa no sudeste. Em 2022, o eixo de maior densidade desloca-se ligeiramente para leste-nordeste, mas sem formar um novo pico independente, sugerindo variação temporária de fluxo em vias adjacentes. Em 2023, o padrão volta a intensificar-se ao longo de toda a faixa central, recuperando a configuração de 2021 e evidenciando que as intervenções pontuais do ano anterior não alteraram a dinâmica estrutural de exposição. Já em 2024, nota-se dispersão moderada em direção ao setor norte, refletida pelo alargamento das zonas amarela e verde, enquanto o máximo global permanece ancorado no mesmo corredor central.

Figura 19 – KDE anual 2021-2024.

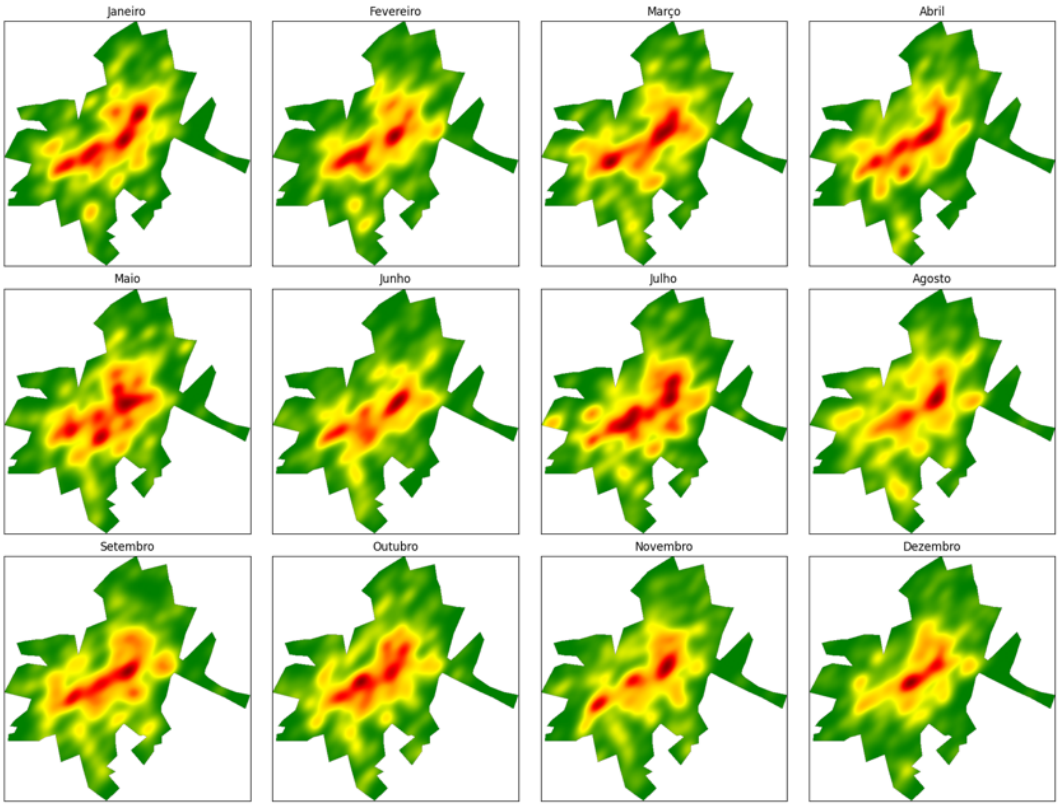


A inspeção da distribuição mensal da densidade permite identificar não apenas a estabilidade do *hotspot* central ao longo dos meses, mas também variações sazonais sutis nas áreas periféricas. Entre janeiro (493 registros; 7,2%) e abril (568; 8,3%), observa-se uma estrutura densa e contínua no corredor central, com ênfase na Avenida Presidente Vargas e adjacências. Em fevereiro (528; 7,7%) e março (622; 9,1%), a intensidade do núcleo nordeste aumenta ligeiramente, indicando possível elevação do fluxo urbano nesse setor, possivelmente associada ao calendário letivo ou atividades comerciais. Maio (631; 9,2%) e junho (593; 8,6%) mantêm a configuração anterior, mas com dispersão mais acentuada para o sudoeste, sugerindo reorganização espacial do risco. Em julho (575; 8,4%) e agosto (635; 9,3%), o *hotspot* central persiste, embora com redução das áreas vermelho-intensas, o que pode estar relacionado a queda temporária de mobilidade urbana no período de férias escolares (Figura 20).

A partir de setembro (553; 8,1%), retoma-se o adensamento central, com maior extensão das zonas de risco elevado, visível nas manchas alaranjadas e vermelhas do setor central-sudoeste. Outubro (548; 8,0%) e novembro (587; 8,6%) mantêm essa estrutura,

apresentando maior coerência interna das áreas críticas. Em dezembro (524; 7,6%), a mancha de risco volta a espalhar ligeiramente para o quadrante leste, mas o núcleo denso permanece fixado no eixo principal, o que reforça a consistência estrutural do padrão de concentração de sinistros na cidade.

Figura 20 – KDE mensal (acumulado 2021-2024).

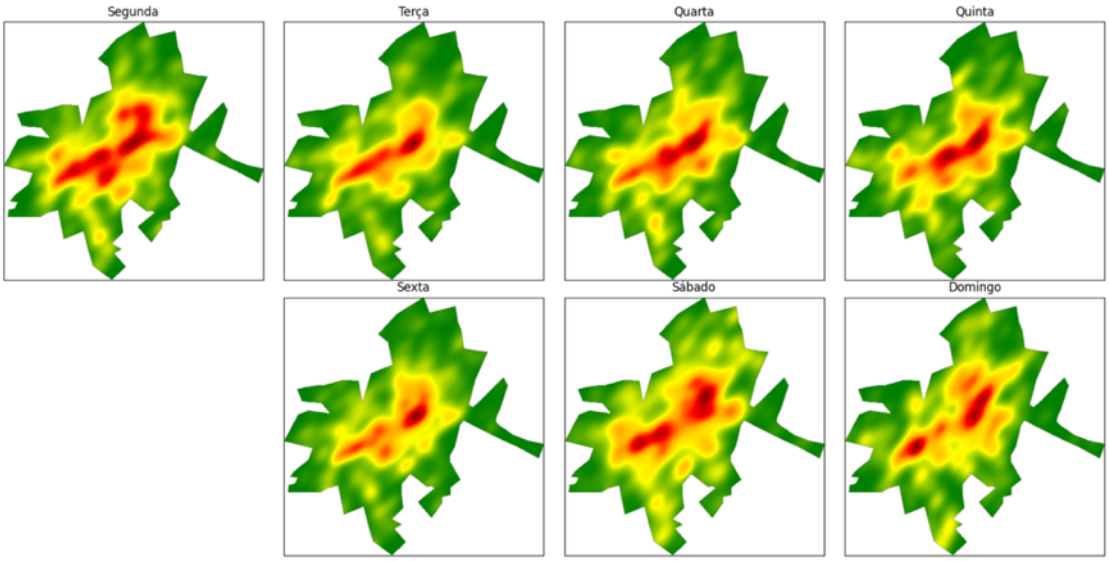


Em relação à distribuição ao longo dos dias da semana, entre segunda e sexta-feira, observa-se um padrão concentrado e linear, fortemente ancorado no eixo principal. A densidade é mais pronunciada nas quartas (1.092 registros; 15,9%) e quintas-feiras (1.070; 15,6%), com ampliação do núcleo central e ramificações para vias secundárias adjacentes, sugerindo sobrecarga típica dos dias de maior atividade comercial e administrativa. Na segunda-feira (1.094; 16,0%), o foco desloca-se ligeiramente para o sudoeste, enquanto a sexta-feira (1.095; 16,0%) apresenta atenuação da densidade no centro e surgimento de focos laterais, o que pode refletir redistribuição de tráfego no encerramento da semana útil.

Nos finais de semana, o padrão altera substancialmente. Aos sábados (880 registros; 12,8%), a densidade permanece relativamente concentrada, mas com menor intensidade, enquanto aos domingos (535; 7,8%) observa-se fragmentação da mancha de calor, com surgimento de múltiplos subnúcleos nos quadrantes norte, leste e sul. A

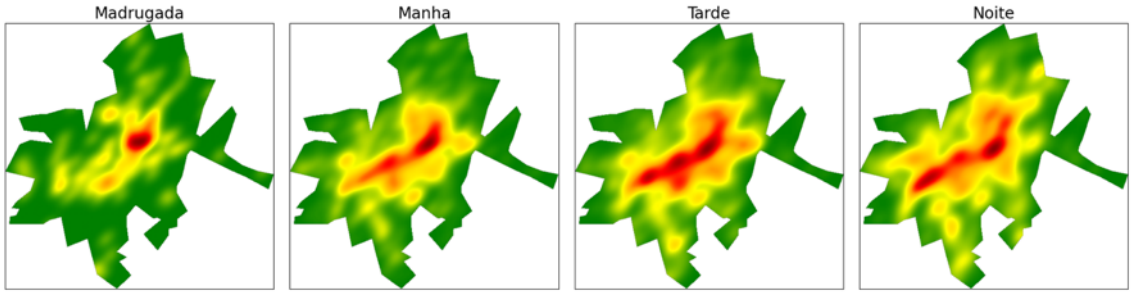
diluição do *hotspot* central nesses dias aponta para uma mobilidade mais difusa, vinculada a deslocamentos não laborais (Figura 21).

Figura 21 – KDE por dia da semana (acumulado 2021-2024).



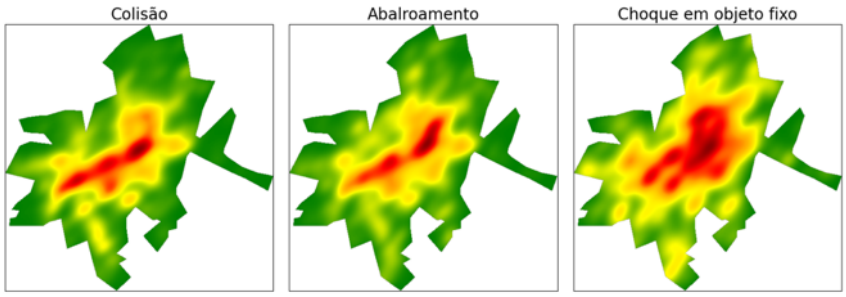
Em relação aos turnos ao longo do dia (Figura 22), o período vespertino concentra 2.951 ocorrências (43,0%), seguido da manhã com 2.151 registros (31,4%), noite com 1.609 (23,5%) e madrugada com 146 (2,1%). Durante a tarde, o mapa evidencia o maior adensamento espacial. O *hotspot* central expande lateralmente e atinge maior continuidade, cobrindo praticamente toda a extensão da Avenida Presidente Vargas e dos eixos transversais conectados à BR-452/GO-174. A sobreposição de fluxos escolares, comerciais e industriais nesse período pode explicar essa configuração. Destaca-se o turno da noite, em que se mantém a predominância do eixo central, mas com aumento de dispersão para quadrantes residenciais periféricos. A redução do volume total de registros nesse período é compensada por uma fragmentação do risco. Já na madrugada, observa-se um padrão pontual e concentrado: embora represente apenas 2,1% dos sinistros no total, há um foco denso no quadrante nordeste do centro expandido.

Figura 22 – KDE por turno (acumulado 2021-2024).



Quando se observa a natureza (Figura 23), a distribuição é apresentada da seguinte forma: colisão (3.062 registros; 44,7%), abalroamento (2.616; 38,2%), choque em objeto fixo (953; 13,9%), outro (160; 2,3%), atropelamento (44; 0,6%), tombamento (11; 0,2%), capotamento (7; 0,1%) e atropelamento de animal (4; 0,1%). Colisão e abalroamento, as naturezas mais frequentes, apresentam manchas densas e contínuas centradas na Avenida Presidente Vargas e seus prolongamentos. Os padrões são semelhantes, com maior alongamento no sentido leste-oeste no caso dos abalroamentos. O grupo choque em objeto fixo exibe uma distribuição mais pulverizada, com manchas intensas em pontos pontuais do quadrante norte e nas extremidades dos eixos viários. Esse padrão é compatível com trechos de aceleração ou saída de perímetro urbano, nos quais a perda de controle veicular é mais provável.

Figura 23 – KDE por natureza (acumulado 2021-2024).

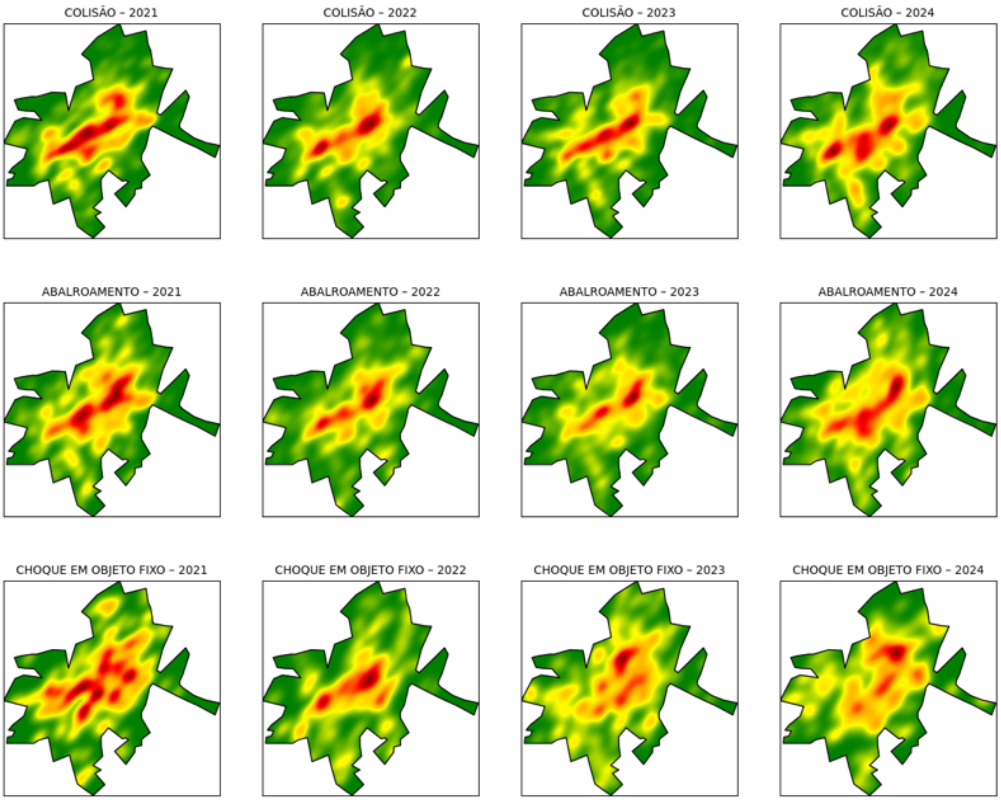


A Figura 24 apresenta a distribuição espaço-temporal das três naturezas de sinistro mais frequentes: colisão, abalroamento e choque em objeto fixo. As colisões mantêm padrão estável ao longo de todo o período, com *hotspot* bem definido no eixo central. Em termos quantitativos, foram registrados 713 casos em 2021 (23,3%), 736 em 2022 (24,0%), 787 em 2023 (25,7%) e 826 em 2024 (27,0%), demonstrando aumento progressivo e persistência estrutural do risco, independentemente de intervenções pontuais. Nos mapas de abalroamento, observa-se maior dispersão espacial,

especialmente em 2022 e 2024, que concentraram 640 (24,5%) e 685 (26,2%) registros, respectivamente. Essa natureza parece responder com maior sensibilidade a alterações locais na malha viária, nos padrões de fluxo ou na ocupação do solo. Ainda assim, o núcleo central permanece presente em todos os anos, reforçando a recorrência em áreas de maior complexidade geométrica.

Já os choques em objeto fixo exibem padrão mais fragmentado e instável, com múltiplos focos dispersos em bordas do perímetro urbano. Foram contabilizados 218 registros em 2021 (22,9%), 265 em 2022 (27,8%), 245 em 2023 (25,7%) e 225 em 2024 (23,6%), mantendo variações discretas ano a ano. Em 2021 e 2022, esses eventos distribuem-se amplamente, mas a partir de 2023 nota-se consolidação de núcleos no quadrante nordeste, sugerindo associação com zonas de expansão urbana ou trechos viários menos regulados. A comparação entre as três naturezas evidencia distintas dinâmicas territoriais e reforça a necessidade de estratégias específicas de prevenção para cada tipo de ocorrência.

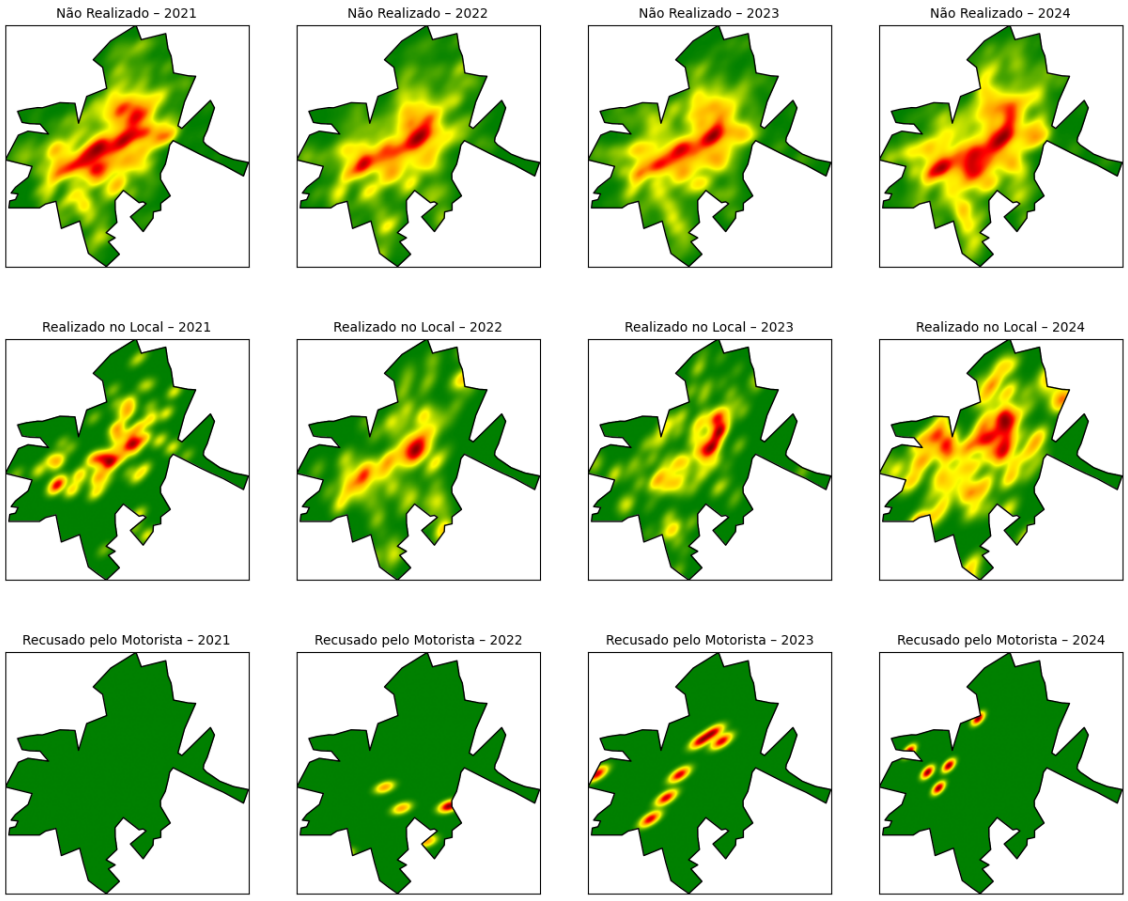
Figura 24 – KDE por natureza (evolução temporal).



Quando observado os testes de alcoolemia, a Figura 25 apresenta a evolução espaço-temporal dos sinistros segundo a situação do exame de alcoolemia entre 2021 e

2024, considerando as três categorias registradas: “não realizado” (12.333 ocorrências), “realizado no local” (816) e “recusado pelo motorista” (18). A categoria predominante, “não realizado”, mantém padrão denso e contínuo ao longo de todo o período, com concentração persistente no eixo central e nas principais vias arteriais, refletindo o comportamento geral da distribuição dos sinistros no município. Nos casos em que o exame foi “realizado no local”, observa-se distribuição mais dispersa e com menor intensidade. Segundo relato de representantes da AMT, a partir de 2022 foi adotado o protocolo de que todo sinistro ocorrido após as 18h deve ser acompanhado da realização de exame de alcoolemia no local, o que pode explicar o aumento e a difusão espacial desse tipo de registro nos anos seguintes. A categoria “recusado pelo motorista” apresenta manchas pontuais e isoladas, distribuídas de forma fragmentada entre 2022 e 2024.

Figura 25 – KDE por exame de alcoolemia (evolução temporal).



Indice de Moran Global (I) e Local (LISA)

A autocorrelação espacial global foi estimada a partir de uma malha hexagonal regular de 250 m, agregando a contagem de sinistros por célula.

A definição dessa dimensão baseou-se em testes exploratórios com diferentes larguras, variando de 100 m a 500 m, avaliando-se o impacto sobre a estabilidade dos indicadores e a identificação de aglomerados. O valor de 250 m demonstrou ser mais adequado ao contexto urbano de Rio Verde por equilibrar detalhamento espacial e robustez estatística, evitando tanto a fragmentação excessiva quanto a diluição de padrões locais relevantes. A matriz de vizinhança utilizada foi de contiguidade de primeira ordem (queen), com pesos binários padronizados por linha, atribuindo peso zero às unidades sem vizinhos ("ilhas"). A significância dos resultados foi verificada por meio de 999 permutações de Monte Carlo, assegurando a detecção de estruturas espaciais consistentes e compatíveis com a escala operacional de planejamento urbano e fiscalização no município.

O resultado obtido foi I = 0,5897 com p = 0,001, rejeitando a hipótese nula de ausência de autocorrelação espacial ao nível de 5%. O valor positivo e elevado indica agrupamento de áreas com altas e baixas frequências de sinistros em posições adjacentes, evidenciando dependência espacial na distribuição dos eventos. A existência de 48 células "ilha" reduziu marginalmente o valor de I, mas não comprometeu a robustez do achado.

Considerando a significância da autocorrelação global, procedeu-se à decomposição local por meio do indicador LISA, a fim de identificar a localização específica dos agrupamentos de sinistros. Essa análise complementa o valor agregado do índice global e permite distinguir áreas críticas e zonas de segurança relativa dentro do tecido urbano. A Tabela 2 apresenta a descrição técnica dos clusters espaciais identificados pelo LISA, classificando-os em Alto-Alto, Baixo-Baixo, Alto-Baixo e Baixo-Alto. Esses agrupamentos representam, respectivamente, áreas críticas de risco, zonas de segurança relativa, pontos atípicos e locais protegidos em zonas de risco.

Os resultados obtidos confirmam que os sinistros ocorrem em corredores viários específicos, coerente com estudos que revelam comportamento semelhante em cidades de porte médio (Plug, Xia e Caulfield, 2011; Shafabakhsh, Famili e Bahadori, 2017). A dependência espacial global justifica a aplicação de indicadores locais de associação (LISA) para detalhar a localização dos *clusters* e apoiar intervenções orientadas pela evidência, conforme recomendado por ANSELIN (1995) e GETIS (2007).

A Figura 26 apresenta os resultados da decomposição local (LISA), que evidenciou predominância de clusters Alto-Alto no núcleo urbano central , notadamente ao longo dos principais corredores arteriais, e de *clusters* Baixo-Baixo em bairros periféricos e área rural. *Outliers* Alto-Baixo e Baixo-Alto ocorreram pontualmente em

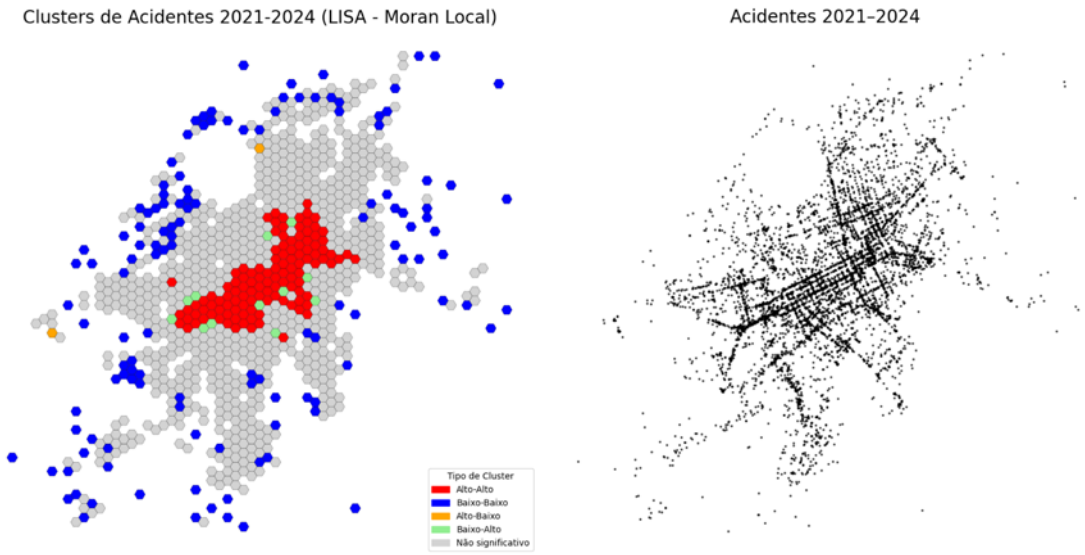
interseções isoladas, enquanto extensas porções da malha foram classificadas como não significativas (p > 0,05), refletindo baixa densidade ou padrão aleatório de sinistros.

Tabela 5 - Descrição dos *Clusters* Espaciais com Base na Autocorrelação Local.

Cor	Tipo de Cluster	Significado técnico
Vermelho	Alto-Alto (HH)	*Hotspots*: hexágonos com alta contagem de sinistros, cercados por outros com alta contagem. São áreas críticas de risco.
Azul	Baixo-Baixo (LL)	*Coldspots*: hexágonos com baixa contagem de sinistros, cercados por outros com baixa contagem. Indicadores de segurança relativa.
Laranja	Alto-Baixo (HL)	*Outlier* negativo: um hexágono com alta contagem, mas cercado por baixa contagem. Pode indicar um ponto atípico, como um cruzamento perigoso isolado.
Verde claro	Baixo-Alto (LH)	*Outlier* positivo: hexágono com baixa contagem, mas cercado por alta contagem. Pode representar um local protegido em uma zona de risco.
Cinza claro	Não significativo	Áreas em que não foi detectada autocorrelação significativa (p > 0,05). O padrão de sinistros podem ser aleatório ou insuficiente para inferência estatística.

Esses resultados reforçam a necessidade de intervenções diferenciadas: ações estruturais nas zonas Alto-Alto para mitigação do risco, monitoramento contínuo dos *outliers* para detecção de mudanças emergentes e abordagem preventiva nas áreas Baixo-Baixo para preservação das condições de segurança observadas.

Figura 26 – *Clusters* (LISA - Moran Local) acompanhados da distribuição espacial dos pontos de sinistros 2021-2024.



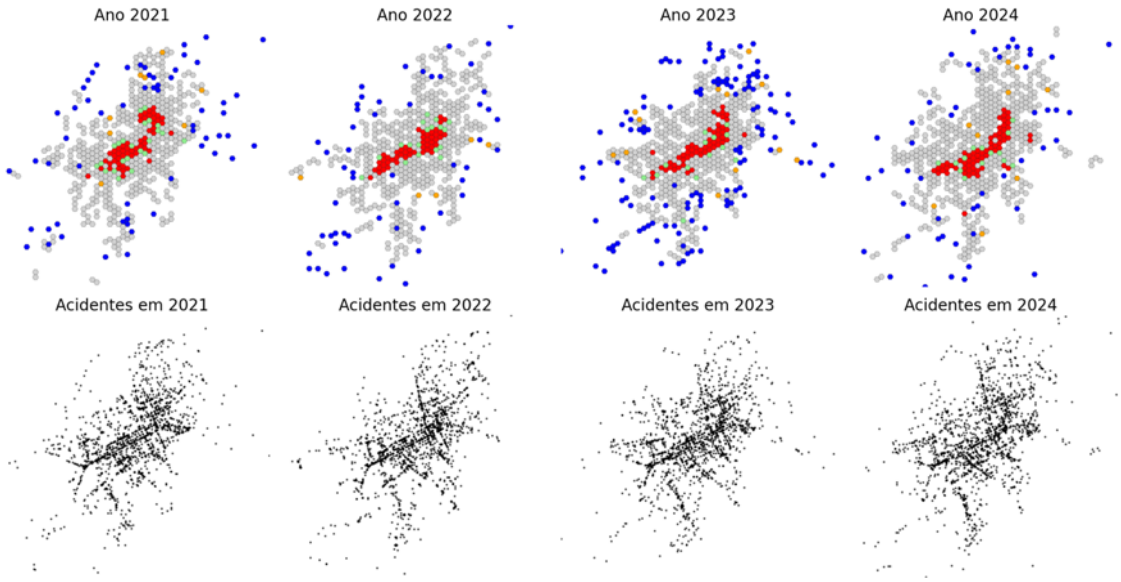
A análise dos indicadores locais de associação espacial (LISA) ao longo dos anos de 2021 a 2024 revela a persistência de um padrão consolidado de autocorrelação espacial

positiva na região central de Rio Verde. Em todos os anos, observa-se a recorrência de agrupamentos classificados como Alto-Alto (HH) ao longo do principal eixo viário leste-oeste, notadamente sobre a Avenida Presidente Vargas e suas extensões. Esses *clusters* significativos mantêm disposição linear, conectando setores centrais e bairros com elevada densidade de tráfego.

A Figura 27 apresenta a evolução temporal dos clusters LISA entre 2021 e 2024, evidenciando a persistência de autocorrelação positiva na região central de Rio Verde.A comparação das quatro séries temporais indica que, embora haja variações na magnitude e na dispersão dos eventos, a estrutura espacial do risco viário permanece relativamente estável no núcleo urbano. Esse padrão é reforçado pelo surgimento consistente de células HH nos mesmos trechos ao longo dos anos, o que reforça a tese de que determinadas áreas da cidade concentram condições crônicas de risco, associadas à configuração geométrica da via, intensidade de tráfego, e ausência de intervenções estruturais de segurança.

Em contrapartida, nas zonas periféricas verifica-se aumento da quantidade de células classificadas como “ilhas” especialmente a partir de 2022. Esse fenômeno sugere expansão do tecido urbano para regiões com pouca ou nenhuma ocorrência registrada.

Figura 27 – Evolução temporal dos *clusters* (LISA - Moran Local) acompanhados da distribuição espacial dos pontos de sinistros.



Esses achados apontam para a existência de uma estrutura espacial recorrente de risco no município, que tende a se concentrar em regiões de maior complexidade viária e

atividade urbana consolidada. A persistência dos *clusters* Alto-Alto nos mesmos setores ao longo dos quatro anos reforça a necessidade de intervenções focalizadas, ancoradas em evidências empíricas, especialmente nas zonas centrais de convergência modal e tráfego misto.

As estimativas de densidade *kernel* e a análise LISA convergem ao evidenciar a concentração de sinistros na Avenida Presidente Vargas. Em virtude dessa recorrência, a investigação foi refinada para a escala de segmento, abrangendo os 7,23 km da via. Esse recorte isolou 353 sinistros depurados (2021-2024) e possibilitou avaliar a distribuição linear dos eventos.

Análise da Avenida Presidente Vargas

A Avenida Presidente Vargas é a espinha dorsal da mobilidade urbana em Rio Verde. Possui extensão de aproximadamente 7,23 km, ligando o bairro Santo Antônio de Lisboa ao entroncamento com as rodovias BR-060 e GO-174 (IBGE, 2020). A via é classificada como arterial, cuja função principal é permitir maior fluidez no fluxo de veículos, com velocidade máxima regulamentada de 50 km/h (PREFEITURA MUNICIPAL DE RIO VERDE, 2023).

Para extrair apenas os pontos de sinistro que permeiam esta avenida, aplicou-se uma expressão regular para extrair latitude e longitude, gerando um GeoDataFrame no sistema WGS 84 (EPSG 4326). Em seguida, os dados foram reprojetados para a projeção *Web Mercator* (EPSG 3857), compatível com mosaicos cartográficos on-line e recomendada para visualizações web.

Para excluir registros indevidamente georreferenciados, primeiramente definiu-se um retângulo de recorte que acompanhava o traçado aproximado da via. Esse *bounding box* cobriu uma faixa média de 120 m ao redor da avenida, eliminando pontos distantes do eixo. Reconhecendo, todavia, que a delimitação retangular poderia manter ruídos, aplicou-se o algoritmo DBSCAN, empregado para detecção de agrupamentos de densidade e remoção de *outliers* espaciais (ESTER *et al*., 1996). A parametrização adotada ($\varepsilon = 250$ m e min_samples = 12) preservou o *cluster* predominante e descartou agrupamentos residuais.

Depois da fase de densidade, construiu-se uma linha central da avenida a partir de quatro vértices ao longo do eixo, convertendo-a em um envelope de 50 m por meio de operação buffer. Apenas os pontos contidos nesse polígono foram retidos, garantindo

correspondência topológica com o logradouro e corrigindo os registros cujas coordenadas divergiam do endereço textual.

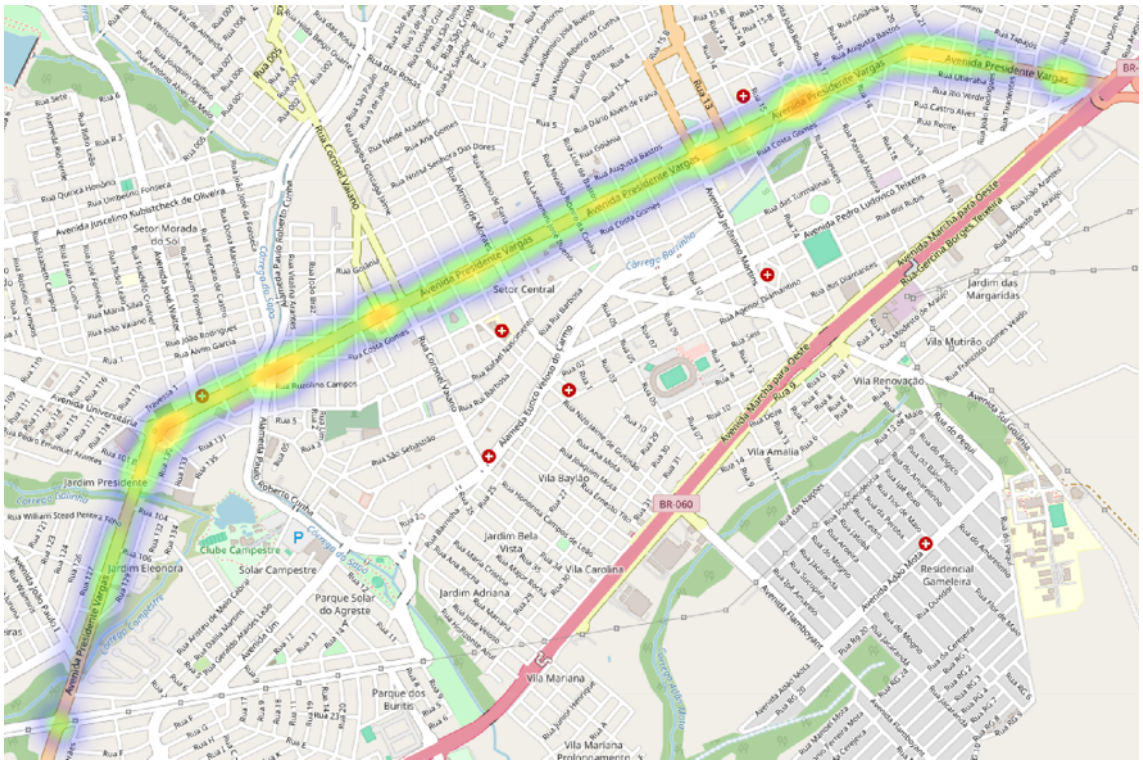
A aplicação combinada dos filtros geográficos, do algoritmo DBSCAN e do buffer linear resultou em um subconjunto de 353 sinistros entre 2021 e 2024. A composição por natureza é apresentada na Tabela 6.

Tabela 6 - Registros na Avenida P. Vargas por natureza.

Natureza	2021	2022	2023	2024	Total
Colisão	63	39	51	65	218
Abalroamento	23	20	19	33	95
Choque em objeto fixo	9	4	7	7	27
Outro	0	1	3	5	9
Atropelamento	1	0	0	2	3
Atropelamento animal	0	0	1	0	1
Total	**96**	**64**	**81**	**112**	**353**

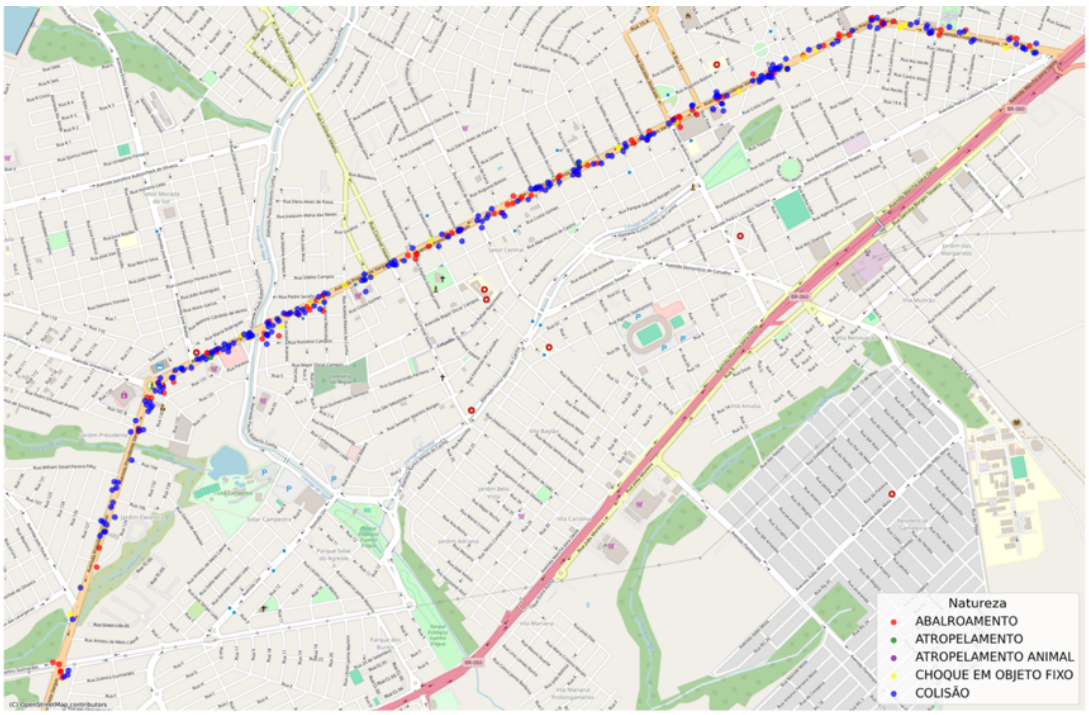
Para representação cartográfica, geraram duas visualizações. No mapa de calor, (Figura 28) estimou a densidade espacial por KDE com largura de banda de 150m (SILVERMAN, 1986). A matriz resultante evidencia gradientes de densidade ao longo da via, evidenciando pontos críticos ao longo dos 7,23 km analisados. Os sinistros classificados na categoria outros foram removidos da análise visual.

Figura 28 - Mapa de calor - Avenida Presidente Vargas (2021 -2024)



Complementarmente, o mapa de pontos confirmou a linearidade da distribuição dos 353 sinistros georreferenciados, após o processo de depuração espacial. Os pontos alinhados ao longo do eixo demonstram aderência ao *buffer* linear de 50 m estabelecido.

Figura 29 - Sinistros na Avenida Presidente Vargas (2021 -2024) por natureza.

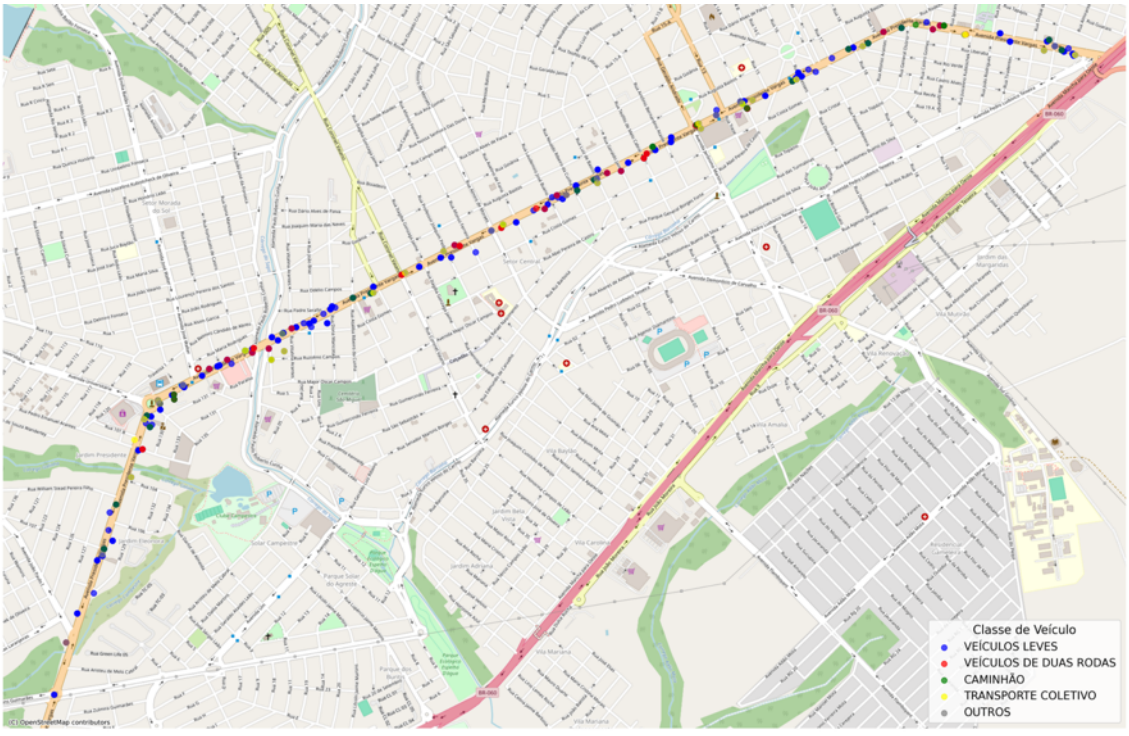


A classificação dos veículos envolvidos nos 353 sinistros registrados na Avenida Presidente Vargas entre 2021 e 2024 foi organizada em cinco grandes categorias: veículos leves, veículos de duas rodas, caminhão, transporte coletivo e outros. Esse agrupamento considerou todos os veículos envolvidos, uma vez que cada sinistro pode envolver múltiplos veículos, totalizando 716 veículos envolvidos.

A categoria veículos leves, que inclui automóveis, camionetas, caminhonetes e utilitários, concentrou 606 registros (85%), correspondendo à ampla maioria dos veículos envolvidos. A segunda classe com maior incidência foi a de veículos de duas rodas, que abrange motocicletas, motonetas, ciclomotores e bicicletas, totalizando 43 ocorrências (6%). A classe caminhão aparece em 30 registros (4%), seguido do grupo transporte coletivo (ônibus e micro-ônibus) com 24 registros (3%) e outros, composta por tipos residuais com 13 registros (2%).

A representação espacial por classe de veículo é evidenciada na Figura 30, que reforça os padrões descritos, revelando que, embora os veículos leves estejam presentes ao longo de toda a avenida, os veículos pesados e coletivos concentram-se com maior densidade nos extremos, próximos às conexões com rodovias e nos polos de maior movimento urbano.

Figura 30 - Sinistros na Avenida Presidente Vargas (2021 -2024) por tipo de veículo.



4.4. Conclusão

Este estudo analisou a evolução temporal e a configuração espacial dos 6.857 sinistros de trânsito sem vítimas registrados em Rio Verde (GO) entre 2021 e 2024, utilizando estimativas de densidade por *kernel* (KDE), índices de autocorrelação espacial (Moran Global e LISA) e o algoritmo DBSCAN para segmentação linear.

A estimativa de densidade por *kernel* evidenciou persistente aglomeração de ocorrências no corredor formado pela Avenida Presidente Vargas e pelas rodovias BR-452/GO-174, padrão que se manteve estável ao longo do quadriênio e foi reforçado pela presença de gradientes de densidade decrescente nas zonas periféricas do município.

A análise de autocorrelação espacial indicou dependência positiva significativa ($I = 0,5897$; $p = 0,001$), consolidando a existência de *clusters* estatisticamente consistentes. A decomposição local (LISA) permitiu identificar células Alto-Alto coincidentes com a malha arterial central e, sobretudo, com a Avenida Presidente Vargas, evidenciando *hotspots* crônicos em trechos de elevada complexidade viária. Para esse eixo, o uso combinado de filtros geográficos e DBSCAN selecionou 353 sinistros distribuídos linearmente ao longo de 7,23 km, confirmando a adequação do método para recortes de segmento viário.

Os achados possuem implicações diretas para a agenda local de segurança viária. A concentração recorrente de sinistros nos principais corredores urbanos corrobora a literatura especializada, que indica que vias com maior hierarquia funcional, como arteriais e rodovias urbanas, tendem a concentrar maior volume de tráfego e velocidades operacionais mais altas, fatores associados ao aumento da frequência e da gravidade dos sinistros (PLUG *et al.,* 2011; XIE; YAN, 2013).

Nesse contexto, evidencia-se a necessidade de intervenções de engenharia em cruzamentos críticos, gestão de velocidade e fiscalização focada em veículos leves, responsáveis por 85% dos registros na Avenida Presidente Vargas. Tais ações convergem com as metas do PNATRANS e da Década 2021-2030, que preconizam reduções graduais de mortes e lesões por meio de infraestrutura segura, fiscalização e educação.

Limitações importantes devem ser reconhecidas. A base analisada restringe-se a sinistros sem vítimas registrados pela AMT, podendo subestimar lesões ou óbitos não reportados. Adicionalmente, o processo de pré-processamento resultou na exclusão de aproximadamente 11% dos registros originais, após etapas de pré-processamento e

limpeza. Embora tais etapas tenham sido necessárias para assegurar a qualidade e a consistência das análises, esse descarte pode acentuar a subnotificação já existente nos sistemas oficiais de registro e, consequentemente, introduzir discrepâncias entre os resultados obtidos e o cenário real. Por fim, inconsistências residuais de georreferenciamento também podem influenciar a precisão espacial, ainda que os procedimentos de limpeza e depuração empregados tenham mitigado esse risco.

Recomenda-se, como continuidade, incorporar sinistros com vítimas e dados de velocidade média veicular, ampliar a série histórica para aferir tendências pós-2024 e aplicar modelos preditivos que considerem intervenções planejadas. Ademais, a adoção de painéis geoespaciais dinâmicos pela gestão municipal poderá viabilizar monitoramento em tempo quase real, potencializando decisões baseadas em evidências e contribuindo para que Rio Verde avance rumo às metas de redução de sinistros estabelecidas em âmbito nacional e internacional.

4.5. Referências Bibliográficas

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Global status report on road safety 2023**. Genebra, 2023.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **Resolução A/RES/74/299: Improving global road safety.** Nova Iorque, 2020.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 10697:2018 – **Elaboração de relatórios estatísticos e operacionais de sinistros de trânsito**. Rio de Janeiro, 2018.

PLUG, C.; XIA, J.; CAULFIELD, C. **Spatial and temporal visualisation techniques for crash analysis.** Accident Analysis and Prevention, v. 43, p. 1937-1946, 2011.

SHAFABAKHSH, G. A.; FAMILI, A.; BAHADORI, M. S. **GIS-based spatial analysis of urban traffic accidents.** Journal of Traffic and Transportation Engineering, v. 4, n. 3, p. 290-299, 2017.

Munasinghe, D. **Spatial Analysis of Urban Road Traffic Accidents Using GIS**. British Journal of Multidisciplinary and Advanced Studies, v. 4, n. 6, p. 70–83, 2023. DOI: 10.37745/bjmas.2022.0368.

MELO, Willian Augusto de; MENDONÇA, Renata Rodrigues. **Caraterização e distribuição espacial dos acidentes de trânsito não fatais**. Cadernos de Saúde Coletiva, Rio de Janeiro, v. 29, n. 1, e010364, 2021.

PAIXÃO, Lúcia Maria Miana Mattos et al. **Acidentes de trânsito em Belo Horizonte: o que revelam três diferentes fontes de informações, 2008 a 2010**. Revista Brasileira de Epidemiologia, São Paulo, v. 18, n. 1, p. 100-114, 2015.

SILVA, Danilo Alves da; PEREIRA, Ruth Bernardes de Lima; ALVES, Marta Maria Malheiros. **Análise sociodemográfica e espacial dos acidentes de trânsito com vítimas fatais em Palmas, Tocantins**. Ciências Biológicas e da Saúde: Pesquisas Básicas e Aplicadas 2. Rio Branco: Stricto Sensu Editora, 2021. p. 269–284.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Estimativas da população 2022**. Rio de Janeiro, 2022. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-da-populacao.htm. Acesso em: 03 jun. 2025.

SERVIÇO FEDERAL DE PROCESSAMENTO DE DADOS. RENAEST – **Relatório Estatístico de Sinistros de Trânsito: Rio Verde, GO (2021-2024)**. Brasília: SENATRAN, 2024. Disponível em: https://www.gov.br/transportes/pt-br/assuntos/transito/arquivos-senatran/docs/renaest. Acesso em: 03 jun. 2025.

MONTELLA, A. **A comparative analysis of hotspot identification methods. Accident Analysis and Prevention**, Oxford, v. 42, p. 571-581, 2010.

XIE, Z.; YAN, J. **Detecting traffic accident clusters with network kernel density estimation and local spatial autocorrelation.** Accident Analysis and Prevention, Oxford, v. 50, p. 477-486, 2013.

ÇALIŞKAN, M.; ANBAROĞLU, B. **Space Time Cube Analytics in QGIS and Python for Hot Spot Detection**. SoftwareX, v. 24, p. 101498, 2023.

ALSAHFI, T. **Spatial and Temporal Analysis of Road Traffic Accidents in Major Californian Cities Using a Geographic Information System**. ISPRS International Journal of Geo-Information, v. 13, p. 157, 2024.

CHAINEY, S.; RATCLIFFE, J. **GIS and Crime Mapping**. Chichester: Wiley, 2013.

O'SULLIVAN, D.; UNWIN, D. **Geographic Information Analysis**. Hoboken: Wiley, 2003.

FOTHERINGHAM, A. S.; BRUNSDON, C.; CHARLTON, M. **Quantitative Geography: Perspectives on Spatial Data Analysis**. London: Sage, 2000.

SILVERMAN, B. W. **Density Estimation for Statistics and Data Analysis**. London: Chapman and Hall, 1986.

SHARIAT-MOHAYMANY, Mohammad; SADAT-KHADEM, Reza; FAGHRI, Mohammad Reza. **Identifying accident hotspots by GIS-based kernel density estimation**. International Journal of Crashworthiness, v. 18, n. 3, p. 292–298, 2013.

ANSELIN, L. **Local indicators of spatial association—LISA**. Geographical Analysis, v. 27, n. 2, p. 93-115, 1995.

GETIS, A. **Reflections on spatial autocorrelation**. Regional Science and Urban Economics, v. 37, p. 491–496, 2007.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Rio Verde: dados gerais**. Rio de Janeiro: IBGE, 2020. Disponível em: https://biblioteca.ibge.gov.br/biblioteca-catalogo.html?id=449182&view=detalhes. Acesso em: 04 ago. 2025.

PREFEITURA MUNICIPAL DE RIO VERDE. **Vias de Rio Verde e suas velocidades máximas permitidas**. Rio Verde, 24 jan. 2023a. Disponível em: https://www.rioverde.go.gov.br/vias-de-rio-verde-e-suas-velocidades-maximas-permitidas/. Acesso em: 10 ago. 2025.

ESTER, M. et al. **A density-based algorithm for discovering clusters in large spatial databases with noise**. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland: AAAI Press, 1996. p. 226–231.

PRASANNAKUMAR, V. et al. Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment. Procedia – Social and Behavioral Sciences, v. 21, p. 317-325, 2011.

5. CAPÍTULO III

ANÁLISE DE DADOS DE SINISTROS DE TRÂNSITO: APLICAÇÃO DE *MACHINE LEARNING* COM CORREÇÃO DE *DATA LEAKAGE*

RESUMO

Este estudo avaliou a viabilidade de construir modelos preditivos confiáveis para sinistros de trânsito registrados em Rio Verde-GO entre 2021 e 2023, empregando um processo rigoroso de detecção e correção de *data leakage*. A base inicial continha 4.601 ocorrências e 18 variáveis, com aumento para 44 variáveis com tratamento inicial; após a eliminação de 26 atributos comprometidos por vazamento de dados, restaram 18 preditores legitimamente disponíveis no momento da decisão. Aplicaram-se técnicas de engenharia de atributos, balanceamento de classes e validação cruzada estratificada (3-*fold*) resultando em 11 modelos supervisionados considerados confiáveis, complementados por análises não supervisionadas que identificaram cinco perfis distintos de sinistros. As métricas de desempenho dos modelos confiáveis variaram de 67,8% a 96,7% de acurácia ou R^2, destacando-se predições robustas para condições da via (seca ou molhada) e combinações de período do dia com finais de semana. A análise de importância de variáveis apontou predominância de fatores geográficos, sazonais e de infraestrutura sobre marcadores estritamente temporais. Além disso, o agrupamento não supervisionado por *K-Means* identificou cinco perfis distintos de sinistros, sendo o *cluster* vespertino responsável por 34,7% das ocorrências. Os resultados confirmam que o controle sistemático de *data leakage* é determinante para a generalização dos modelos e demonstram o potencial do aprendizado de máquina como ferramenta de apoio às políticas de segurança viária em municípios de porte médio.

Palavras-chave: sinistros de trânsito; aprendizado de máquina; *data leakage*; modelagem preditiva; Rio Verde-GO.

ABSTRACT

This study assessed the feasibility of constructing reliable predictive models for traffic crashes recorded in Rio Verde-GO between 2021 and 2023, employing a rigorous process for detecting and correcting data leakage. The original dataset contained 4,601 records and 18 variables, later expanded to 44 variables after initial processing; following the exclusion of 26 attributes compromised by leakage, 18 predictors legitimately available at the decision-making moment were retained. Feature engineering, class balancing, and stratified 3-fold cross-validation techniques were applied, resulting in 11 supervised models considered reliable, complemented by unsupervised analyses that identified five distinct accident profiles. The performance metrics of the reliable models ranged from 67.8% to 96.7% accuracy or R^2, with robust predictions observed for road surface conditions (dry or wet) and combinations of time of day and weekends. Variable importance analysis indicated a predominance of geographic, seasonal, and infrastructure-related factors over strictly temporal markers. Furthermore, unsupervised clustering via K-Means identified five distinct accident profiles, with the afternoon cluster accounting for 34.7% of all incidents. The results confirm that systematic control of data leakage is critical to ensuring model generalization and demonstrate the potential of

machine learning as a decision-support tool for traffic safety policies in medium-sized cities.

Key words: traffic crashes; machine learning; data leakage; predictive modeling; Rio Verde-GO.

5.1. Introdução

A análise preditiva de sinistros de trânsito tornou-se instrumento estratégico para a gestão da segurança viária e o planejamento urbano, permitindo antecipar cenários de risco, otimizar a alocação de recursos de emergência e fundamentar políticas públicas baseadas em evidências. O crescimento do tráfego motorizado em centros de porte médio, como Rio Verde, GO, reforça a demanda por métodos que descrevam padrões de ocorrência e subsidiem intervenções pontuais em vias e bairros críticos.

Nesse contexto, o aprendizado de máquina oferece um conjunto de técnicas capazes de ajustar funções preditivas diretamente a partir dos dados, reduzindo a imposição de pressupostos paramétricos característicos da estatística clássica (JORDAN; MITCHELL, 2015). Algoritmos supervisionados, como *Random Forest, Extra Trees, Regressão Logística, Naive Bayes, XGBoost e LightGBM*, já demonstraram utilidade na classificação de gravidade e na estimativa de condições adversas, enquanto abordagens não supervisionadas permitem identificar estruturas latentes úteis à segmentação de perfis de risco.

A construção de modelos confiáveis, entretanto, enfrenta o desafio recorrente do *data leakage*, fenômeno em que variáveis indisponíveis no momento da decisão são inadvertidamente incorporadas ao treinamento, produzindo métricas infladas e modelos inviáveis em produção (KAUFMAN *et al.,* 2012; KAPOOR; NARAYANAN, 2023). Em bases de dados de sinistros de trânsito, o vazamento manifesta-se principalmente por atributos temporais diretos ou derivados, estados sazonais e identificadores de alta cardinalidade que possibilitam memorizar casos específicos em vez de aprender padrões generalizáveis. A literatura indica que tais formas sutis de vazamento permanecem subdetectadas em muitos estudos aplicados, comprometendo a validade externa dos resultados.

Apesar da pertinência dessa discussão, são escassas as investigações empíricas que combinem detecção sistemática de *data leakage* com a avaliação de modelos de aprendizado de máquina alimentados exclusivamente por variáveis legítimas em cenários locais. O conjunto de 4.601 registros de sinistros ocorridos em Rio Verde entre 2021 e 2023 oferece oportunidade ímpar para examinar, de forma controlada, a influência desse problema metodológico sobre a capacidade preditiva dos modelos.

Dessa forma, o presente estudo tem por objetivo avaliar a viabilidade de desenvolver modelos preditivos confiáveis para sinistros de trânsito em Rio Verde, GO, ocorridos entre 2021 e 2023, mediante a aplicação de uma metodologia de detecção e correção de *data leakage*, garantindo que apenas variáveis legitimamente disponíveis no momento da decisão integrem o processo de aprendizado.

5.2. Material e Método

5.2.1. Coleta e pré-processamento de dados

A base de dados foi constituída a partir de planilha contendo informações sobre sinistros sem vítimas, encaminhada pela Agência Municipal de Mobilidade e Trânsito (AMT), de Rio Verde, - GO, com 4.601 registros compreendidos entre 01 de janeiro de 2021 e 31 de dezembro de 2023. Os registros contemplam dados sobre o momento do sinistro, como dia da semana e período do dia; dados sobre o local, incluindo bairro, tipo de via e condições climáticas; e dados sobre os condutores e veículos, como categoria da CNH, teste de alcoolemia, quantidade de veículos envolvidos e tipo de avarias constatadas.

O pré-processamento dos dados foi realizado no intuito de garantir a consistência estrutural e a viabilidade analítica do conjunto de registros de sinistros. O *dataset* original (Quadro 2), composto por 18 variáveis, apresentava inconsistências como dados ausentes e tipos de dados mistos. Além disso, apresentava características que impunham desafios à modelagem, como o desbalanceamento entre categorias e a presença de variáveis com alta cardinalidade.

Quadro 2 - Base de dados original.

Variável	Descrição
Dia Semana	Dia da semana em que ocorreu o sinistro
Período	Período do dia em que ocorreu o sinistro
Zona	Classificação da zona em que ocorreu o sinistro
Natureza	Tipo de sinistro ocorrido
Bairro	Nome do bairro em que ocorreu o sinistro
Controle Tráfego	Indicação de tráfego no local
Pista	Descrição das características da pista em que ocorreu o sinistro

Pavimento	Tipo de pavimento da via
Condições da via	Descrição da via em que ocorreu o sinistro
Condições do tempo	Condições climáticas no momento do sinistro
Mês	Mês em que ocorreu o sinistro
Ano	Ano em que ocorreu o sinistro
Habilitação	Situação de habilitação do condutor envolvido
Categoria	Categoria da CNH do condutor
Teste Álcool	Realização do exame toxicológico
Valor (Mg/L)	Valor do exame toxicológico
Veículos Envolvidos	Tipo de veículo do(s) envolvido(s)
Avarias	Classificação do dano/avaria nos veículos

A variável "HORA", por exemplo, exigiu correção de heterogeneidade nos formatos, assegurando extração do período do dia, sendo um procedimento essencial para análises temporais subsequentes. Além disso, variáveis demográficas como "IDADE" foram categorizadas em faixas etárias, conforme recomendações da literatura especializada (McCartt *et al*., 2009), permitindo maior sensibilidade na detecção de padrões epidemiológicos. O tratamento resultou em adição de 26 variáveis (Quadro 3) ampliando a base de dados para 44 variáveis, após esta transformação.

Quadro 3 – Variáveis adicionadas.

Variável	**Descrição**
hora_numerica	Hora do dia em formato numérico (0 a 23)
periodo_detalhado	Período detalhado do dia (madrugada, manhã, tarde, noite)
horario_pico	Indicador binário de ocorrência em horário de pico (Sim/Não)
hora_sin	Transformação cíclica da hora: seno(2π·hora/24)
hora_cos	Transformação cíclica da hora: cosseno(2π·hora/24)
faixa_etaria	Faixa etária do envolvido (18-25, 26-35, 36-50, 51-65, 65+)
ano	Ano do sinistro em formato numérico (2021, 2022, 2023)
mes	Mês do sinistro em formato numérico (1 a 12)
dia_semana	Dia da semana em formato numérico (0–6, sendo domingo=0)
dia_mes	Dia do mês em que ocorreu o sinistro (1 a 31)
semana_ano	Número da semana no ano
eh_feriado	Indicador binário de feriado (0 = não feriado, 1 = feriado)
eh_fim_semana	Indicador binário para fim de semana (0 = não, 1 = sim)
eh_inicio_mes	Indicador binário para início de mês (0 = não, 1 = sim)
eh_fim_mes	Indicador binário para final de mês (0 = não, 1 = sim)
trimestre	Trimestre do ano (1 a 4)
estacao	Estação do ano correspondente à data (Verão, Outono, Inverno, Primavera)
mes_sin	Transformação cíclica do mês: seno(2π·mês/12)
mes_cos	Transformação cíclica do mês: cosseno(2π·mês/12)
dia_semana_sin	Transformação cíclica do dia da semana: seno(2π·dia/7)
dia_semana_cos	Transformação cíclica do dia da semana: cosseno(2π·dia/7)

zona_periodo	Variável de interação Zona × Período (ex.: Urbana_Matutino)
condicoes_via	Condição da via padronizada (bom_asfalto, chuva_asfalto, etc.)
periodo_fds	Interação entre período do dia e fim de semana (ex.: manhãa_fds, noite_semana).
pico_fds	Interação entre pico e fim de semana (ex.: pico_fds, normal_semana)
estacao_periodo	Interação entre estação do ano e período do dia (ex.: verao_manha).

Outro desafio metodológico consistiu no desbalanceamento nas categorias de natureza dos sinistros, comprometendo a robustez estatística para tarefas de classificação. As categorias com frequência inferior a 1% foram consolidadas em grupos mais amplos (Fernández *et al*., 2018) com base na gravidade e no mecanismo de lesão. Adicionalmente, a variável "bairro", com 247 categorias únicas, foi simplificada por meio da retenção dos 25 bairros mais incidentes, responsáveis por 68% dos casos. Os demais foram agrupados como "outros_bairros", estratégia que preserva 85% da informação discriminativa e segue o princípio de Pareto aplicado à distribuição espacial (Micci-Barreca, 2001; Newman, 2005).

A engenharia de variáveis foi implementada de forma a capturar dimensões temporais e contextuais mais refinadas. Foram extraídas informações como período do dia, horários de pico, dia da semana, mês, trimestre e estações do ano, além de indicadores binários de datas específicas. Para contemplar a natureza cíclica do tempo, transformações trigonométricas foram aplicadas às variáveis temporais. Adicionalmente, foram criadas também variáveis de interação, tais como "zona_periodo", para representar sinergias entre fatores espaciais e temporais. A etapa final consistiu na padronização dos tipos de dados, no tratamento direcionado de valores ausentes e na validação cruzada da coerência entre variáveis.

5.2.2. Machine Learning

O aprendizado de máquina compreende um conjunto de técnicas estatísticas que extraem padrões de bases de dados para produzir inferências e previsões automatizadas. Diferente da modelagem estatística clássica, na qual a estrutura funcional é imposta a priori, os algoritmos de aprendizado de máquina ajustam parâmetros diretamente a partir dos dados, reduzindo suposições paramétricas (Jordan; Mitchell, 2015). Essa abordagem alcança duas tarefas essenciais na análise de sinistros de trânsito: (i) reconhecimento de estruturas latentes, sendo útil na identificação de perfis de risco, e (ii) estimação de

funções preditivas para variáveis categóricas ou contínuas. A literatura consolida o campo em paradigmas supervisionado e não supervisionado, cada qual associado a hipóteses distintas sobre a disponibilidade de rótulos nos dados (Bishop, 2006).

Algoritmos não supervisionados

O aprendizado não supervisionado parte do pressuposto de que não existem rótulos pré-definidos e, portanto, o objetivo é revelar estruturas latentes nos dados. Nos problemas de agrupamento, as observações são particionadas segundo critérios de similaridade: o *K-Means* minimiza a soma das dispersões intraclasse ao redor de centroides (MACQUEEN, 1967); o DBSCAN identifica regiões de alta densidade, admite formatos de *cluster* arbitrários e reconhece ruídos (ESTER *et al*., 1996); os métodos hierárquicos, a exemplo da ligação de Ward, geram dendrogramas que descrevem relações de proximidade em múltiplas escalas (WARD, 1963). Para redução de dimensionalidade, a Análise de Componentes Principais transforma variáveis correlacionadas em eixos ortogonais que retêm a maior proporção de variância explicada (JOLLIFFE, 2002). Na detecção de anomalias, o *Isolation Forest* aplica o princípio de aleatoriedade florestal para isolar instâncias em menor número de divisões, caracterizando pontos atípicos (LIU; TING; ZHOU, 2008). Tais procedimentos possibilitam segmentar vias com padrões homogêneos de sinistros, revelar zonas críticas e investigar *outlier*s espaço-temporais sem recorrer a variáveis-alvo.

Algoritmos supervisionados

No aprendizado supervisionado, parte-se de amostras rotuladas para estimar funções que mapeiam atributos explanatórios em respostas categóricas ou contínuas. O algoritmo *Random Forest*, por exemplo, constrói múltiplas árvores de decisão sobre subconjuntos de dados e atributos, agregando previsões para reduzir variância (BREIMAN, 2001). O algoritmo *Extra Trees* estende essa ideia ao introduzir divisão totalmente aleatória dos pontos de corte, diminuindo correlação entre árvores e acelerando o ajuste (GEURTS; ERNST; WEHENKEL, 2006). A Regressão Logística, enquadrada nos modelos lineares generalizados, estima a probabilidade de cada classe por meio da função logística e serve de base para análise de riscos por ser interpretável em termos de *odds* ratio (HOSMER; LEMESHOW; STURDIVANT, 2013). O algoritmo *Naive Bayes*, fundamentado no teorema de Bayes e na suposição de independência condicional entre preditores, oferece classificação rápida, especialmente útil em cenários

de alta dimensionalidade (MURPHY, 2012). Já os métodos de gradiente reforçado avançaram com o *XGBoost*, que integra regularização e estimação de segundas derivadas para otimização eficiente (CHEN; GUESTRIN, 2016), e com o *LightGBM*, que adota partições por histogramas e crescimento folha-a-folha para lidar com grandes volumes de dados de maneira escalável (KE *et al*., 2017). No contexto de sinistros de trânsito, esses algoritmos permitem classificar períodos críticos, prever gravidade e quantificar incertezas prognósticas com validação cruzada estratificada.

Data Leakage

Data leakage, ou vazamento de dados, constitui-se um dos problemas mais críticos em aprendizado de máquina, ocorrendo quando informações que não estariam disponíveis no momento da predição em um cenário real são inadvertidamente incluídas durante o treinamento do modelo. Este fenômeno resulta em modelos com performance artificialmente inflada durante a validação, mas que falham completamente quando aplicados a dados prospectivos em ambiente de produção.

O problema de *data leakage* é particularmente prevalente em aplicações de análise temporal e estudos observacionais, em que a distinção entre informação preditiva legítima e vazamento de informação futura pode ser sutil. Conforme documentado por KAUFMAN *et al*. (2012) e KAPOOR & NARAYANAN (2023), uma proporção significativa de estudos em aprendizado de máquina aplicado pode estar comprometida por formas não detectadas de vazamento de dados, resultando em conclusões científicas e aplicações práticas fundamentalmente falhas.

O vazamento de dados pode manifestar-se de diversas formas: através da inclusão de variáveis que são consequência direta do evento que se pretende prever, da utilização de informações futuras para prever eventos passados, da incorporação de identificadores únicos que permitem ao modelo memorizar casos específicos ao invés de aprender padrões generalizáveis, ou através de proxies que indiretamente contêm a informação que se pretende prever.

No contexto deste estudo, as técnicas supervisionadas foram aplicadas às tarefas de classificação e regressão relacionadas a variáveis de interesse operacional, como gravidade, tipo de via e condições da pista. Já as técnicas não supervisionadas foram utilizadas para identificar estruturas latentes e perfis de risco, posteriormente incorporados como atributos adicionais nos modelos supervisionados.

5.2.3. Modelagem e Validação

O processo de modelagem foi precedido por uma etapa de refinamento destinada à detecção e correção de *data leakage*. Essa etapa assegurou que apenas variáveis legítimas, disponíveis no momento da decisão, fossem utilizadas no treinamento e na validação dos modelos. O controle sistemático do vazamento de dados constituiu, portanto, um filtro metodológico essencial antes da aplicação dos algoritmos supervisionados e não supervisionados, evitando métricas infladas e garantindo a validade externa dos resultados

Na sequência, a modelagem contemplou a implementação de seis algoritmos com pressupostos distintos: *Random Forest, Extra Trees, Logistic Regression, Naive Bayes, XGBoost* e *LightGBM*. Esses algoritmos foram aplicados a 11 variáveis-alvo (targets) definidas no estudo, resultando em 11 modelos supervisionados considerados confiáveis após validação cruzada estratificada (3-*fold*). Além disso, técnicas não supervisionadas, como *K-Means* e PCA, foram empregadas para identificar cinco perfis de sinistros. Esses clusters não foram tratados como modelos independentes, mas incorporados como variáveis explicativas adicionais nos modelos supervisionados, constituindo uma estratégia de engenharia de atributos que ampliou a capacidade preditiva.

A reprodutibilidade foi garantida através da fixação de *seeds* aleatórias (valor 42) em todos os pontos críticos do *pipeline*, incluindo divisão de dados, inicialização de algoritmos e processos de amostragem. A validação foi conduzida utilizando 25% dos dados como conjunto de teste, mantido completamente separado durante o treinamento.

O processo de modelagem foi concebido para assegurar a robustez metodológica e a reprodutibilidade dos resultados, combinando diferentes paradigmas de aprendizado de máquina e múltiplos algoritmos em paralelo. A escolha por aplicar abordagens supervisionadas e não supervisionadas fundamenta-se na necessidade de, por um lado, estimar funções preditivas para variáveis de interesse operacional e, por outro, identificar padrões latentes não rotulados que enriquecessem a compreensão do fenômeno.

Nos modelos supervisionados, o conjunto de 4.601 registros foi dividido em 75% para treino (≈ 3.451 registros) e 25% para teste (≈ 1.150 registros), mantendo estratificação nas tarefas de classificação. A reprodutibilidade foi garantida por meio da fixação de *seeds* aleatórias e pela utilização de validação cruzada estratificada (3-*fold*). Foram implementados seis algoritmos de referência com pressupostos distintos: *Random*

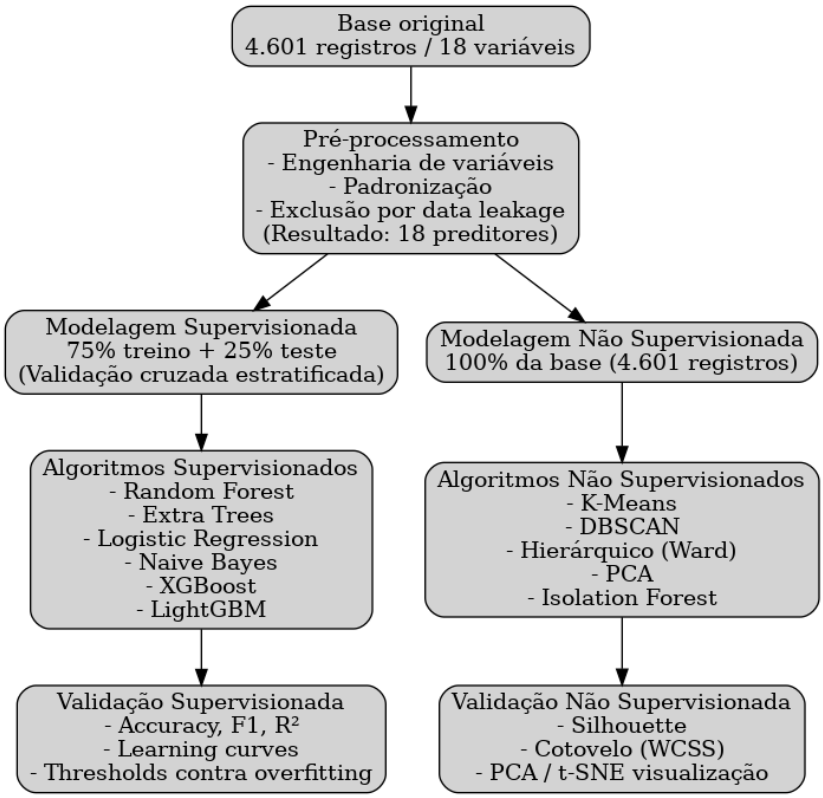
Forest, Extra Trees, Logistic Regression, Naive Bayes, XGBoost e LightGBM. Essa diversidade metodológica teve como objetivo reduzir o viés decorrente da dependência em um único modelo e avaliar o desempenho sob diferentes estruturas de ajuste estatístico e computacional.

Nos modelos não supervisionados, optou-se por utilizar a totalidade dos registros (100%), uma vez que não havia rótulos a serem previstos. Técnicas de agrupamento (*K-Means*, DBSCAN, hierárquico de *Ward*), redução de dimensionalidade (PCA) e detecção de anomalias (*Isolation Forest*) foram aplicadas para identificar perfis de risco, padrões espaço-temporais e *outliers* relevantes. Os *clusters* derivados não foram tratados como variáveis-alvo, mas incorporados como features adicionais em modelos supervisionados, seguindo prática consolidada de engenharia de atributos.

De modo geral, as aplicações de aprendizado de máquina foram conduzidas adotando inicialmente os parâmetros padrão da biblioteca de implementação, com ajustes experimentais de hiperparâmetros a partir da acurácia obtida em validação cruzada. Para mitigar riscos de sobreajuste e de métricas inflacionadas por *data leakage*, foi implementado um sistema de validação em múltiplas camadas. Esse processo incluiu: (i) análise de correlação entre variáveis remanescentes e *proxies* temporais; (ii) sinalização automática de modelos com *accuracy* > 0,98 ou $R^2 > 0,95$; e (iii) inspeção de *learning curves* para monitorar o equilíbrio entre treino e validação. Apenas os modelos que cumpriram esses critérios foram considerados confiáveis, compondo o conjunto final de predições.

A Figura 31 apresenta o fluxograma metodológico da modelagem e validação, sintetizando as etapas de pré-processamento, divisão de bases, aplicação dos algoritmos supervisionados e não supervisionados, e procedimentos de validação adotados.

Figura 31 – Fluxograma metodológico.



O fluxograma sintetiza a articulação entre as etapas metodológicas e os resultados alcançados. O pré-processamento e a correção de *data leakage* atuaram como filtros, garantindo a consistência das variáveis utilizadas. A modelagem supervisionada gerou predições aplicáveis a condições da via, gravidade e padrões temporais, enquanto a análise não supervisionada produziu os cinco *clusters* que caracterizaram perfis distintos de sinistros. Esses blocos se integram no refinamento metodológico que fundamentou os nove modelos confiáveis reportados nos resultados, articulando a função de cada técnica ao objetivo final da pesquisa.

5.3. Resultados e Discussão

5.3.1. Correção de *Data Leakage*

O algoritmo utilizado na implementação de detecção de vazamento de dados identificou variáveis que apresentavam correlações artificialmente perfeitas ou

mapeamentos determinísticos que comprometeriam a validade dos modelos em aplicações práticas.

A análise revelou múltiplas categorias de vazamento de dados presentes no *dataset:*

- Vazamento Temporal Direto: A variável "hora_numerica" apresentava mapeamento unívoco com "periodo_detalhado", em que cada faixa horária correspondia deterministicamente a um período específico do dia. As horas 0-5 mapeavam invariavelmente para "Madrugada", 6-11 para "Manhã", 12-17 para "Tarde" e 18-23 para "Noite". Esta relação permitiria a um modelo alcançar 100% de acurácia na predição do período do dia, mas utilizando informação temporal precisa que não estaria disponível no momento da decisão de resposta a um sinistro em um cenário operacional real.
- Vazamento por Variáveis Derivadas Temporais: As variáveis trigonométricas "hora_sin", "hora_cos", "mes_sin", "mes_cos", "dia_semana_sin" e "dia_semana_cos" foram identificadas como fontes de vazamento por serem transformações matemáticas diretas das variáveis temporais originais. Estas variáveis mantinham a capacidade de reconstruir perfeitamente a informação temporal original, perpetuando o problema de vazamento através de uma representação matematicamente diferente, mas informativamente equivalente.
- Vazamento por Variáveis de Interação Temporal: Variáveis compostas como "periodo_fds" (combinação de período do dia com indicador de fim de semana), "pico_fds" (interação entre horário de pico e fim de semana), "estacao_periodo" (combinação de estação do ano com período do dia) e "zona_periodo" (interação entre zona geográfica e período temporal) foram identificadas como fontes indiretas de vazamento. Embora não fossem diretamente temporais, estas variáveis permitiam a inferência de informações temporais específicas através das combinações, mantendo a capacidade de predição artificial.
- Vazamento por Variáveis de Estado Temporal: Indicadores como "eh_fim_semana", "eh_inicio_mes", "eh_fim_mes", "eh_feriado" representavam estados temporais que, embora conceitualmente diferentes

das variáveis de tempo absoluto, mantinham capacidade de inferência temporal. Estas variáveis permitiriam a reconstrução de informações sobre quando o sinistro ocorreu, informação que não estaria disponível no momento da tomada de decisão preventiva.

Nesse sentido, o *dataset* reduziu de 44 para 18 variáveis confiáveis, após remoção de atributos comprometidos por *data leakage*. Este refinamento buscou garantir que os modelos de aprendizado de máquina fossem construídos apenas com informações disponíveis no momento da decisão, respeitando os princípios de validade metodológica e segurança operacional (Kaufman *et al.,* 2012; Kapoor & Narayanan, 2023).

A decisão de remoção de 27 variáveis foi fundamentada na análise de que, em um cenário operacional real, sistemas preditivos de sinistros devem ser baseados exclusivamente em informações disponíveis antes da ocorrência do evento ou em seu momento inicial, quando decisões de resposta ainda podem ser implementadas. Variáveis que permitam a reconstrução de informações temporais específicas violam este princípio fundamental. As variáveis removidas são apresentadas na Tabela 7:

Tabela 7 - Variáveis Removidas por Categoria de *Data Leakage.*

Categoria	Variáveis Removidas	Motivo da Remoção
Temporais Diretas (3)	hora_numerica, periodo_detalhado, Período	Mapeamento determinístico 1:1 entre hora e período do dia.
Temporais Derivadas (7)	horario_pico, hora_sin, hora_cos, mes_sin, mes_cos, dia_semana_sin, dia_semana_cos	Transformações matemáticas que preservam informação temporal original.
Estados Temporais (12)	ano, Ano, mes, dia_semana, dia_mes, semana_ano, eh_feriado, eh_fim_semana, eh_inicio_mes, eh_fim_mes, trimestre, estacao	Indicadores que permitem reconstrução de informações temporais específicas.
Interações Temporais (4)	periodo_fds, pico_fds, estacao_periodo, zona_periodo	Combinações que mantêm capacidade de inferência temporal.
Identificadores (1)	bairro	Código único que permite memorização ao invés de generalização.

A remoção das categorias apresentadas demonstrou como estas variáveis comprometeriam a validade dos modelos em aplicações prospectivas:

- Temporais Diretas: Apresentavam correlação perfeita (r = 1.0) com *outcomes* temporais. Um modelo que sabe que são "14:30" pode prever com 100% de certeza que é "Tarde", mas esta informação não estaria

disponível em um sistema de alerta preventivo que deve identificar condições de risco antes dos sinistros ocorrerem.

- Temporais Derivadas: As transformações trigonométricas (sin/cos) mantinham toda a informação temporal original em formato matematicamente diferente. Um modelo poderia reconstruir o horário exato a partir de hora_sin e hora_cos, perpetuando o problema de vazamento através de uma representação alternativa.
- Estados Temporais: Variáveis como "eh_fim_semana" ou "trimestre" permitiam inferência de contexto temporal específico. Um sistema que sabe que um sinistro ocorreu "no primeiro trimestre de 2023" tem acesso à informação temporal que não estaria disponível para predições prospectivas.
- Interações Temporais: Combinações como "periodo_fds" (manhã + fim de semana) preservavam capacidade de inferência temporal através de relacionamentos compostos. Mesmo sem acesso direto ao horário, o modelo poderia inferir padrões temporais específicos através destas combinações.
- Identificadores: Códigos únicos como "bairro" permitiam ao modelo memorizar casos específicos ao invés de aprender padrões generalizáveis. Um modelo que memoriza que "acidente #1234 no código 567 foi grave" não está aprendendo padrões aplicáveis a novos casos, mas sim memorizando *outcomes* históricos específicos.

Cada uma destas categorias violava o princípio fundamental de que modelos preditivos devem basear-se exclusivamente em informações que estariam legitimamente disponíveis no momento da aplicação prática, seja para prevenção, triagem inicial ou alocação de recursos de emergência.

Validação da Correção de *Data Leakage*

Para garantir a completude da remoção de vazamento de dados, foi implementado um sistema de validação em múltiplas camadas. Primeiro, foi conduzida uma análise de correlação entre todas as variáveis remanescentes e potenciais *proxies* temporais,

verificando que nenhuma variável mantida permitia a reconstrução de informações temporais específicas com correlação superior a 0.3.

A segunda camada consistiu na implementação de um sistema que identificasse modelos com performance com acurácia superior a 99,9% para problemas de classificação ou R^2 superior a 0.98 para problemas de regressão, sendo automaticamente sinalizado para investigação adicional de possível vazamento residual, caso encontrado. Este sistema de detecção baseado em *thresholds* de performance foi fundamental para identificar vazamentos sutis que poderiam escapar à análise de correlação direta.

Na terceira camada, foi conduzida uma análise de *learning curves* para todos os modelos desenvolvidos, verificando que o gap entre performance de treino e validação permanecia nos limites aceitáveis (diferença inferior a 5% para problemas de classificação). Modelos com gaps excessivos foram sinalizados como potencialmente comprometidos por *overfitting*, que pode ser indicativo de vazamento de dados residual.

Os resultados desta validação em múltiplas camadas confirmaram a eliminação efetiva do *data leakage*. Nenhum dos modelos finais apresentou performance suspeita que indicasse vazamento residual, e todos demonstraram gaps apropriados entre treino e validação, confirmando que aprenderam padrões generalizáveis ao invés de memorizar casos específicos.

A redução sistemática de 44 para 18 variáveis, representando diminuição de 60% nas dimensões originais, ilustra a extensão do problema de *data leakage* presente no *dataset* original e ressalta a importância crítica de análises rigorosas de vazamento de dados em aplicações de aprendizado de máquina. Esta magnitude de redução, embora substancial, foi necessária para garantir a integridade metodológica e a aplicabilidade prática dos modelos desenvolvidos.

Após a limpeza de *data leakage*, foi implementado um processo de engenharia de características para compensar a perda de informação temporal e maximizar o potencial preditivo das variáveis restantes. Este processo incluiu a criação de dez variáveis-alvo (targets) derivadas das características originais dos sinistros.

As variáveis-alvo foram desenvolvidas seguindo critérios de relevância prática e aplicabilidade operacional. O target "gravidade_detalhada" classifica sinistros em quatro níveis de severidade baseados na natureza do incidente, enquanto "gravidade_binaria" simplifica esta classificação em sinistros graves versus não graves. O "condicoes_adversas_score" combina informações meteorológicas e de estado da via para criar um índice de risco ambiental.

O processo de codificação de variáveis categóricas foi implementado utilizando *Label Encoding* para variáveis binárias, *One-Hot Encoding* para variáveis de baixa cardinalidade e combinações de *Label Encoding* com *Frequency Encoding* para variáveis de alta cardinalidade. Esta abordagem híbrida otimiza o balance entre expressividade dos dados e eficiência computacional.

A criação de variáveis de agregação envolveu o cálculo de estatísticas descritivas (média, desvio padrão, contagem) agrupadas por categorias geográficas e contextuais. Por exemplo, foram calculadas médias de frequência de sinistros por bairro e dia da semana, permitindo que os modelos capturem padrões locais e temporais específicos.

Variáveis polinomiais foram introduzidas para capturar relações não lineares, particularmente através da aplicação de transformações quadráticas e logarítmicas às variáveis de frequência e densidade. A normalização robusta foi aplicada utilizando *RobustScaler* e *StandardScaler*, garantindo que variáveis com diferentes escalas contribuam equitativamente para os modelos.

5.3.2. Análise exploratória

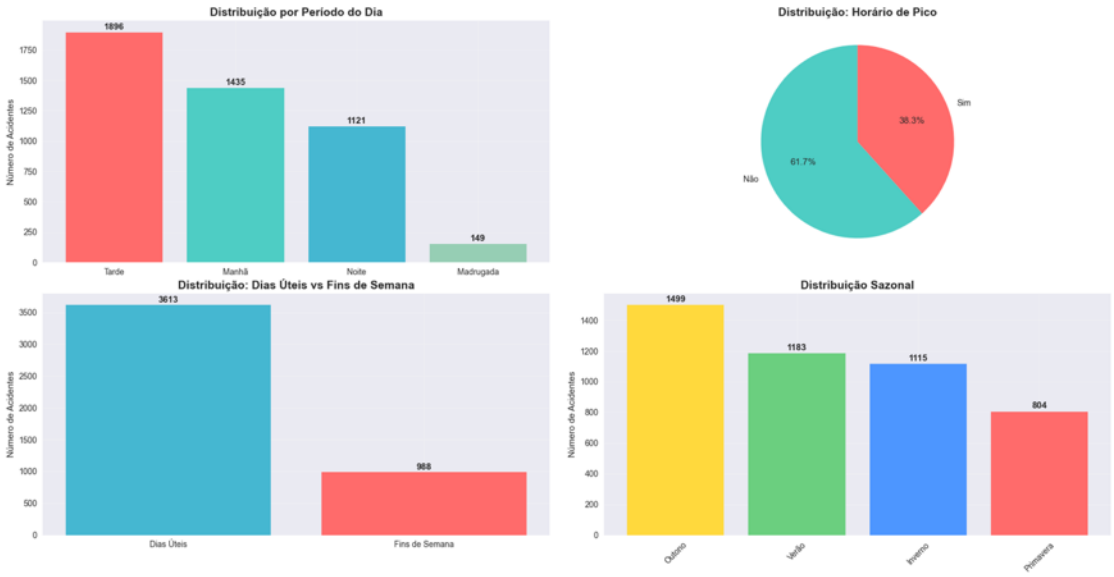
Distribuição espacial

O conjunto de dados é composto de 4.601 registros de sinistros de trânsito, tratados e organizados, conforme os critérios de limpeza, padronização e transformação de variáveis. Após o tratamento dos dados, foram geradas variáveis temporais e categóricas adicionais, possibilitando análises mais refinadas sobre padrões temporais e geográficos dos sinistros. As figuras a seguir representam os principais resultados obtidos.

A análise temporal evidencia que o período vespertino concentra a maior parte dos sinistros, totalizando 1.896 registros, o que representa 41,2% do total. Em seguida, observa-se alta incidência no período da manhã, com 1.435 ocorrências (31,2%), e à noite, com 1.121 (24,4%). Os registros durante a madrugada são significativamente menores, somando apenas 149 casos (3,2%). Além disso, 38,3% dos sinistros ocorreram em horários considerados de pico (entre 7h e 9h e 17h e 19h), o que pode estar relacionado aos momentos de maior movimentação urbana. Também é notável que a maioria dos sinistros ocorre em dias úteis (78,5%), enquanto os fins de semana concentram apenas 21,5% das ocorrências, o que representa uma sub-representação de 7,1 pontos percentuais

em relação ao esperado (considerando que os fins de semana compreendem 2 de 7 dias da semana). Em relação à distribuição sazonal, o outono é a estação com maior frequência de sinistros (32,6%), seguido do verão (25,7%), inverno (24,2%) e primavera (17,5%).

Figura 32 - Distribuição dos sinistros por período do dia, horário de pico, dias úteis versus fins de semana e estações do ano.



A Figura 32 apresenta uma síntese visual dos padrões temporais identificados nos registros de sinistros de trânsito. São exploradas diversas dimensões temporais com o objetivo de elucidar tendências e concentrações ao longo do tempo.

No canto superior esquerdo, a distribuição por hora do dia revela uma concentração de sinistros entre 7h e 20h, com picos notáveis durante os horários de pico matutino e vespertino. A barra em vermelho reforça essas faixas críticas, destacando a necessidade de maior atenção a essas janelas temporais.

Centralizado no topo, o período vespertino é apresentado com o maior número de registros (1.896), seguido pela manhã (1.435) e pela noite (1.121). A madrugada é o período menos crítico, com apenas 149 registros.

À direita, observa-se a distribuição por dia da semana. Os dias úteis mantêm uma distribuição relativamente uniforme, com leve variação entre segunda e sexta-feira, enquanto os finais de semana, especialmente o domingo, apresentam queda acentuada na frequência de sinistros.

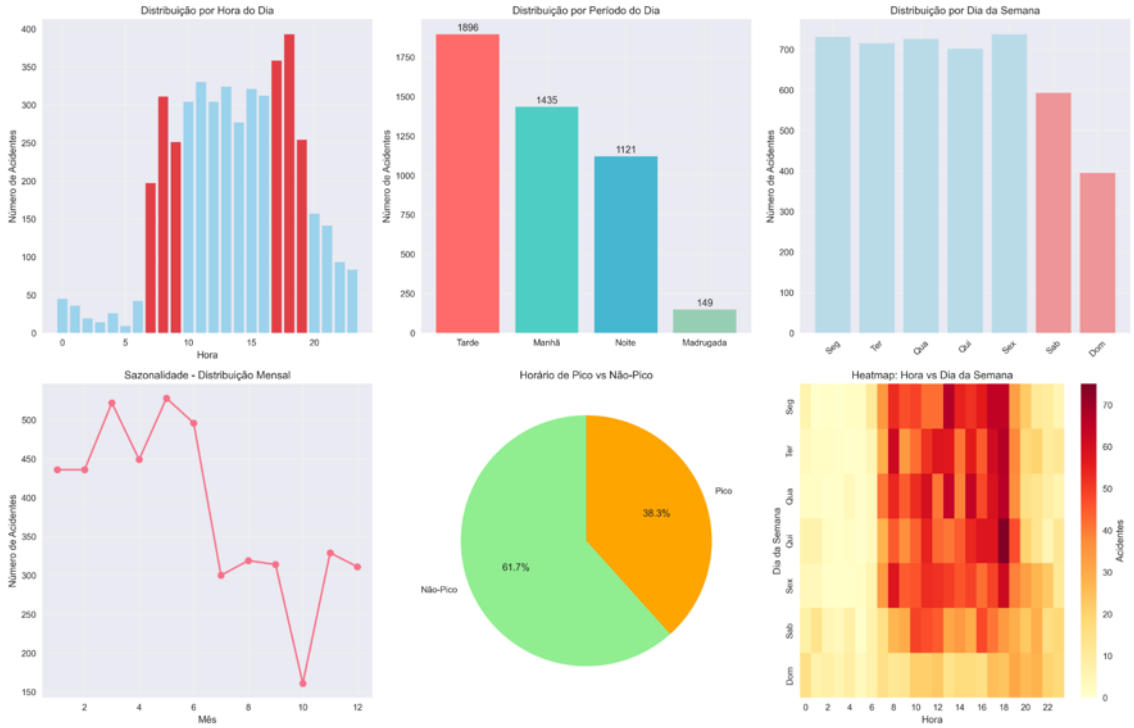
No canto inferior esquerdo, a distribuição mensal evidencia a sazonalidade dos sinistros, com maior incidência nos primeiros seis meses do ano e queda acentuada nos

últimos três meses. O outono e o verão são os períodos sazonais mais críticos, conforme análises complementares.

Ao centro, o gráfico de pizza mostra a divisão entre sinistros ocorridos em horários de pico (38,3%) e fora do pico (61,7%).

No canto inferior direito, o *heatmap* (mapa de calor) correlaciona hora e dia da semana, apontando que os sinistros se concentram entre 11h e 18h nos dias úteis, evidenciando o impacto direto da rotina laboral e do tráfego urbano nas ocorrências.

Figura 33 - Análise dos padrões temporais.



Distribuição geográfica

A Tabela 8 apresenta os 10 bairros com mais registros de sinistro. Destacam a agregação de Outros Bairros (1.277 registros), Setor Central (710), Jardim Goiás (388) e Bairro Popular (289). Juntos, os cinco primeiros bairros representam aproximadamente 62,5% do total de sinistros.

A categoria "Outros_Bairros" agrega localidades menos recorrentes, porém reforça que poucos bairros acumulam a maior parte dos sinistros reportados.

Tabela 8 - Top 10 Bairros com mais sinistros.

Posição	Bairro	Número de Sinistros (%)
1	Outros bairros	1.277 (27.8%)
2	Setor Central	710 (15.4%)
3	Jardim Goiás	388 (8.4%)
4	Bairro Popular	289 (6.3%)
5	Setor Pauzanes	213 (4.6%)
6	Setor Morada do Sol	196 (4.3%)
7	Vila Maria	187 (4.1%)
8	Jardim Presidente	179 (3.9%)
9	Parque Bandeirante	150 (3.3%)
10	Bairro Martins	85 (1.8%)

5.3.3. Aprendizagem de máquina não supervisionado

A presente análise teve por objetivo a identificação de perfis distintos de sinistros de trânsito por meio de técnicas de agrupamento (*clustering*), possibilitando, assim, a compreensão de padrões latentes nos dados e a proposição de estratégias direcionadas de prevenção. Os registros foram organizados em cinco *clusters*, com base em variáveis associadas às condições das vias, período do dia, dia da semana e outros fatores contextuais. A qualidade geral do agrupamento foi avaliada pelo coeficiente de *Silhouette*, cujo valor de 0,122, embora modesto, indica separabilidade moderada entre os grupos identificados.

Análise da Otimização do Número de *Clusters*

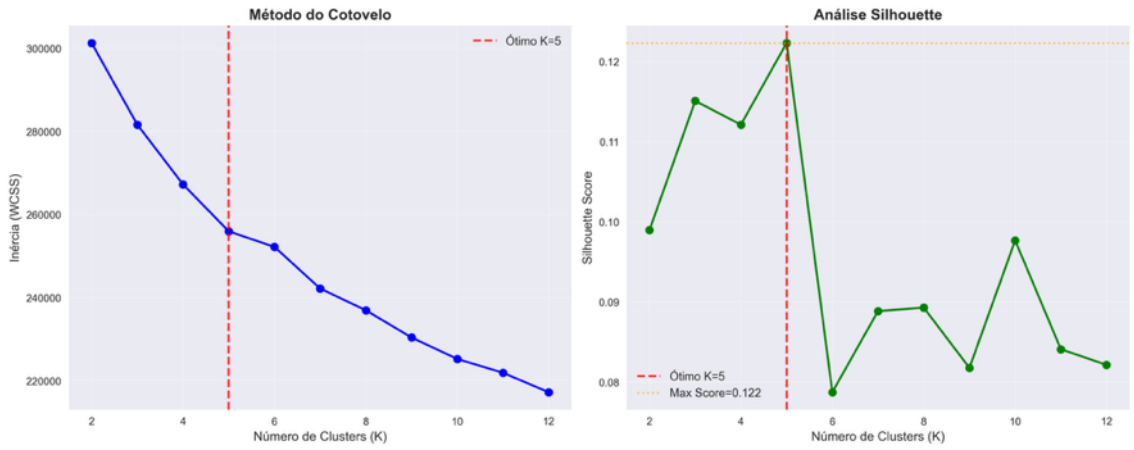
A Figura 33 apresenta dois dos métodos mais utilizados para a definição do número ótimo de *clusters* em uma análise de agrupamento utilizando o algoritmo *K-means*: o Método do Cotovelo (à esquerda) e a Análise do Coeficiente de *Silhouette* (à direita).

No gráfico da esquerda, observa-se o comportamento da inércia (ou soma das distâncias quadradas dentro dos *clusters* – WCSS) à medida que o número de *clusters* (K) aumenta. O "cotovelo" visível no ponto K=5 indica uma inflexão na curva, sugerindo que a adição de mais *clusters* a partir desse ponto resulta em reduções marginais na inércia. Assim, K=5 é visualmente identificado como o valor ótimo.

No gráfico da direita, é apresentada a média do coeficiente de *Silhouette* para diferentes valores de K. Esse coeficiente avalia a qualidade dos agrupamentos, indicando a adequação de cada ponto ao seu respectivo *cluster*. O pico observado em K=5 (com valor máximo de 0.122) reforça a seleção desse valor como ideal, visto que representa o melhor equilíbrio entre coesão *intra-cluster* e separação *inter-cluster*.

A convergência entre os dois métodos em K=5 fornece uma evidência da adequação desse número de agrupamentos, apoiando análises subsequentes com base nesse valor.

Figura 34 - Determinação do Número Ótimo de *Clusters* via Métodos do Cotovelo e Silhouette.



Análise de *Clusters* no Espaço PCA e t-SNE

A representação dos *clusters* gerados a partir do algoritmo de *K-Means*, projetados no espaço bidimensional por meio da Análise de Componentes Principais (PCA) é apresentado na Figura 34. Essa técnica de redução de dimensionalidade permite visualizar os agrupamentos de maneira simplificada, conservando o máximo possível da variância original dos dados. No gráfico, cada ponto representa um sinistro de trânsito, enquanto as cores diferenciam os cinco *clusters* identificados. As cruzes vermelhas marcam os centroides de cada *cluster*, ou seja, os pontos médios que representam o centro geométrico de cada agrupamento.

As componentes principais PC1 e PC2 explicam, respectivamente, 10,4% e 8,1% da variância dos dados, somando uma variância total explicada de 18,5%. Embora a porcentagem de variância explicada pelas duas primeiras componentes não seja alta, a visualização oferece uma perspectiva útil para identificar padrões, sobreposições e distanciamentos entre os agrupamentos.

A disposição espacial dos *cluster*s revela características específicas em termos de proximidade e dispersão. Por exemplo, o *Cluster* 2 (em roxo), localizado à esquerda, apresenta-se bem definido e relativamente separado dos demais, sugerindo um padrão mais homogêneo entre os registros que o compõem. Em contrapartida, os *Clusters* 0, 1 e 3 mostram interseções visuais significativas, sugerindo possível sobreposição de características ou menor distinção entre seus perfis. Já o *Cluster* 4, em amarelo, apresenta uma distribuição mais dispersa, indicando maior heterogeneidade interna.

A visualização no espaço PCA permite avaliar visualmente a coerência dos agrupamentos formados, identificar *clusters* com maior ou menor separabilidade e orientar a interpretação dos padrões comportamentais extraídos nos dados de sinistros.

A Figura 35 apresenta a visualização dos agrupamentos (*clusters*) gerados a partir da técnica t-SNE (t-*Distributed Stochastic Neighbor Embedding*), aplicada aos dados de sinistros de trânsito. Esse método é comumente utilizado para a redução de dimensionalidade e é especialmente eficaz na preservação das relações locais em conjuntos de dados complexos e de alta dimensionalidade.

A dispersão observada entre os *clusters* reforça a heterogeneidade dos dados, evidenciando que os sinistros possuem características particulares que os agrupam de maneira coerente em termos de comportamento e contexto. A técnica de t-SNE permite observar, de forma mais clara, a separação entre os *clusters* que não é facilmente perceptível em análises com múltiplas variáveis simultâneas.

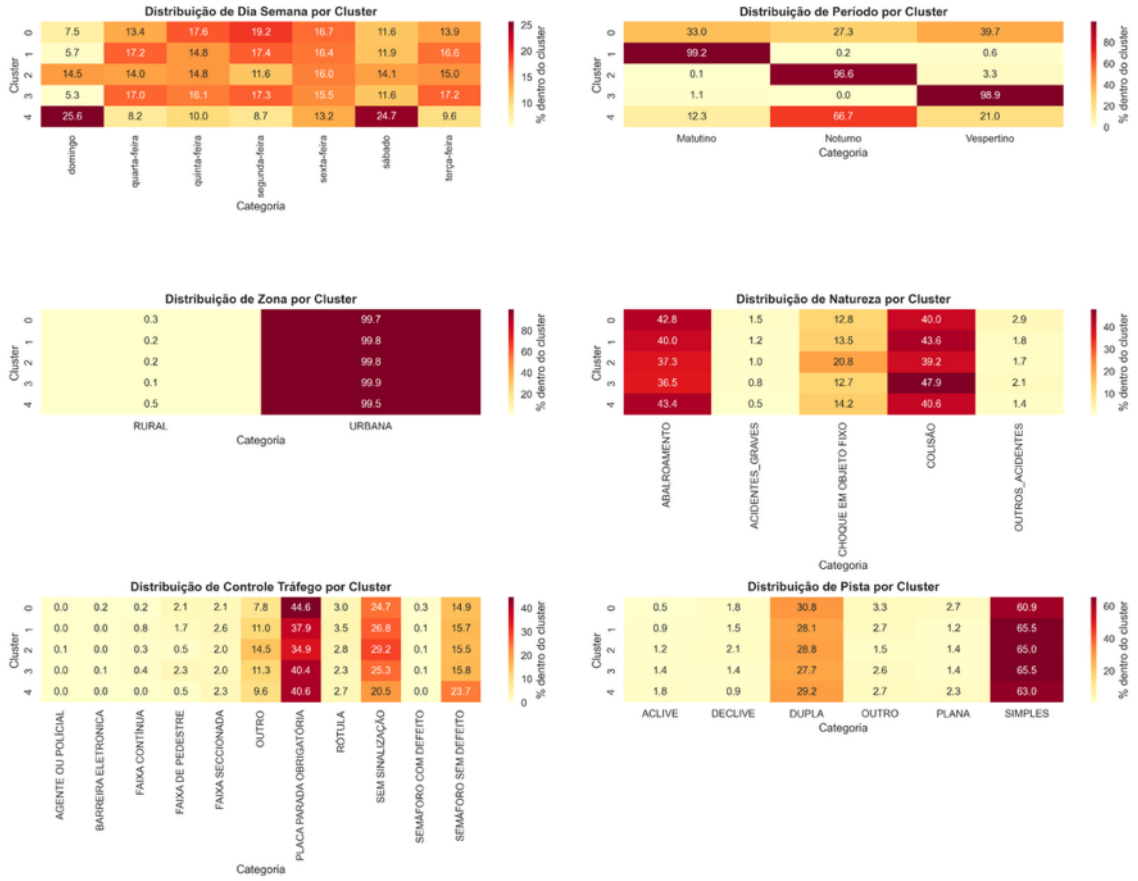
Figura 35 - Visualização dos *Clusters* com PCA.

Figura 36 - Visualização dos *Clusters* com t-SNE.

Seis mapas de calor que ilustram as distribuições relativas das variáveis categóricas mais relevantes entre os *clusters* identificados na análise (Figura 36). No canto superior esquerdo, observa-se a distribuição dos dias da semana, destacando que o *Cluster* 4 concentra 25,6% dos casos aos domingos, sugerindo comportamentos específicos de risco em fins de semana. Ao lado, o mapa de calor referente ao período do dia indica que o *Cluster* 1 é dominado por sinistros matutinos (99,2%), enquanto o *Cluster* 3 concentra-se no período vespertino (98,9%).

A distribuição das naturezas dos sinistros possui predominância de colisão em quase todos os *clusters*, exceto no *Cluster* 2, em que há maior proporção de choque em objeto fixo. Na linha inferior, a distribuição por zona confirma que os sinistros se concentram quase exclusivamente em área urbana em todos os *clusters*. A distribuição por tipo de controle de tráfego evidencia onde semáforo sem defeito e placas de parada obrigatória aparecem frequentemente, mas com variação de intensidade por *cluster*. Por fim, o último gráfico destaca o tipo de pista, evidenciando que a maior parte dos sinistros em todos os grupos ocorreu em pistas simples (acima de 60% em todos os *clusters*).

Figura 37 - Distribuições por *Cluster*.

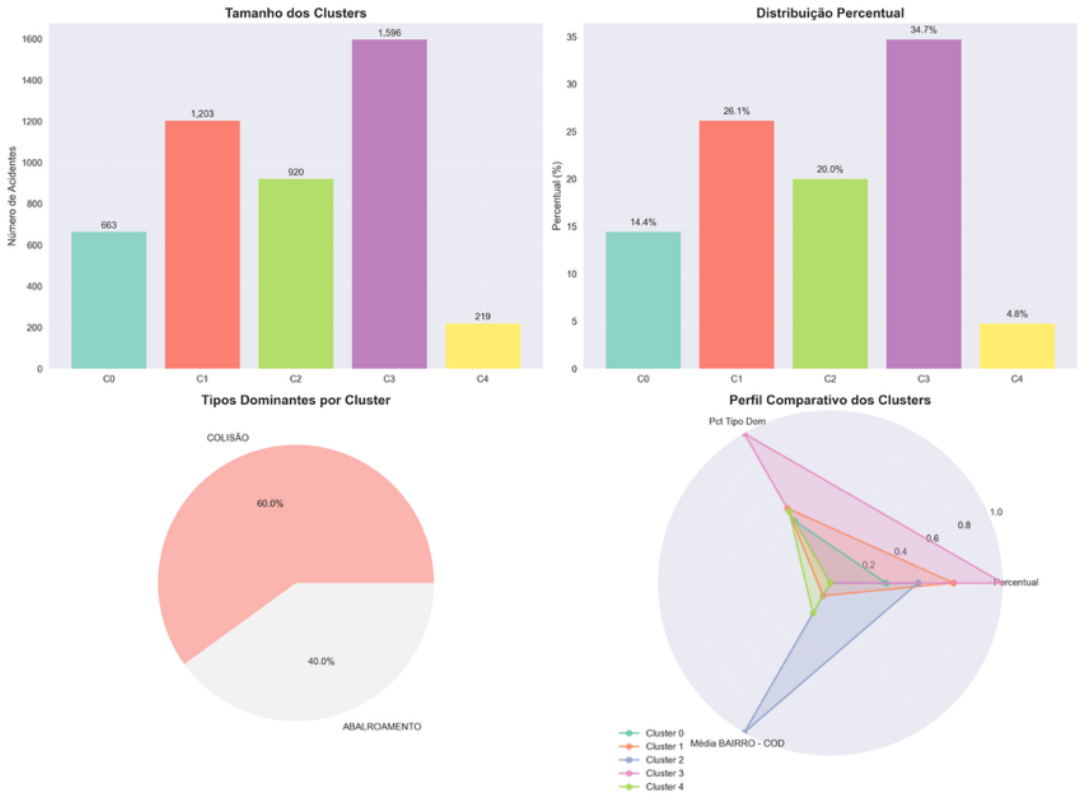


Comparações Estatísticas entre *Clusters*

A Figura 37 apresenta uma série de visualizações que permitem uma análise comparativa entre os *clusters* identificados a partir do agrupamento de dados de sinistros de trânsito. Na parte superior esquerda, o gráfico de barras exibe o tamanho absoluto dos *clusters* em número de sinistros: o Cluster 3 concentra a maior quantidade (1.596), seguido pelo *Cluster* 1 (1.203), *Cluster* 2 (920), *Cluster* 0 (663) e, por fim, o *Cluster* 4 com 219 ocorrências. Já o gráfico superior direito mostra a distribuição percentual de cada cluster em relação ao total, confirmando a dominância relativa do *Cluster* 3, com 34,7% do total de sinistros agrupados.

Na parte inferior esquerda, o gráfico de pizza ilustra a predominância dos tipos de sinistros, destacando que colisões representam 60% dos casos predominantes por *cluster*, seguidas por abalroamentos com 40%. Essa visualização evidencia o tipo de sinistro mais frequente nos agrupamentos realizados. Por fim, o gráfico radar localizado no canto inferior direito permite a comparação simultânea de três atributos dos *clusters*: a média do código do bairro, a proporção do tipo de sinistro dominante (Pct Tipo Dom) e o percentual geral do *cluster*. Essa combinação de atributos é útil para identificar *clusters* com perfis mais homogêneos ou com maior impacto estatístico sobre o conjunto de dados.

Figura 38 - Análise Comparativa da Estrutura e Perfil dos *Clusters* Identificados.



O *Cluster* 0, composto por 663 sinistros (14,4%), apresentou um perfil singular de elevada ocorrência em vias molhadas (55,8% contra 9,5% na média geral), sugerindo influência direta das condições meteorológicas na ocorrência dos sinistros. Este grupo requer atenção específica, sobretudo para monitoramento em períodos de chuva ou pista escorregadia, ainda que a prioridade de intervenção seja considerada baixa.

O *Cluster* 1 (1.203 sinistros; 26,1%) é fortemente associado ao período matutino, com 99,2% das ocorrências concentradas nesse intervalo, frente a 31,7% na amostra geral.

O *Cluster* 2 reúne 920 sinistros (20,0%) e caracteriza-se pela predominância de ocorrências no período noturno (96,6%), bem acima da média geral de 26,5%.

Já o Cluster 3, maior dentre os grupos (1.596 sinistros; 34,7%), apresenta clara associação com o período vespertino, com 98,9% das ocorrências.

Por fim, o *Cluster* 4 (219 sinistros; 4,8%) agrega características peculiares, com 25,6% das ocorrências registradas aos domingos (frente a 8,5% no geral) e 66,7% no período noturno.

Os principais *insights* revelam que o *Cluster* 3 é o mais representativo, enquanto o *Cluster* 4 apresenta especificidades importantes. A distribuição dos agrupamentos, variando de 219 a 1.596 registros, reforça a heterogeneidade dos sinistros e a importância de abordagens segmentadas.

Análise dos padrões temporais por *cluster*

A Figura 38 intitulada 'Distribuição dos Padrões Temporais por *Cluster* de Sinistros' apresenta um conjunto de mapas de calor que descrevem a distribuição temporal dos sinistros de trânsito segundo os *clusters* identificados na análise de agrupamento.

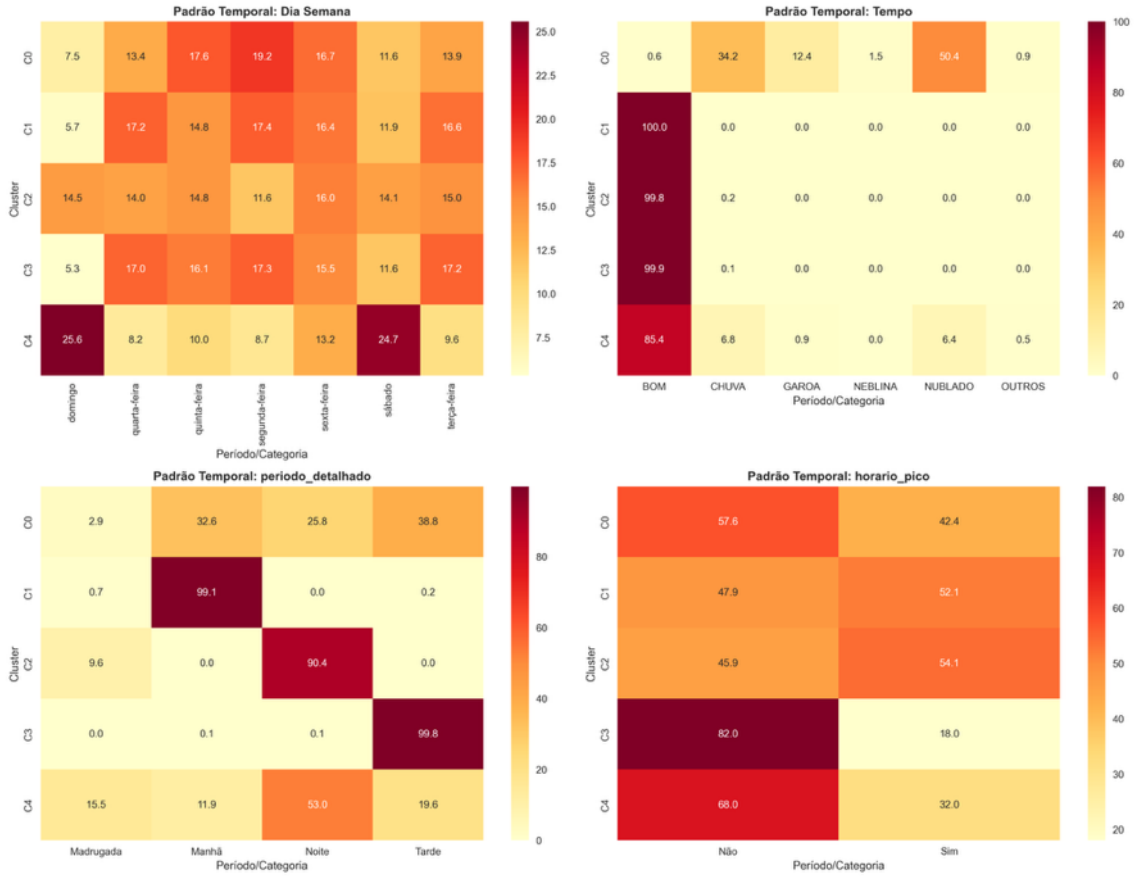
No canto superior esquerdo, a matriz "Padrão Temporal: Dia da Semana" indica a frequência relativa de sinistros por dia da semana em cada *cluster*. Observa-se que o *Cluster* 4 destaca-se significativamente aos domingos (25,6%) e sábados (24,7%), sugerindo que esse grupo concentra sinistros em fins de semana, e pode estar relacionado ao comportamento recreativo dos condutores.

O gráfico superior direito, "Padrão Temporal: Tempo", revela as condições meteorológicas predominantes durante os sinistros. Os *clusters* 1, 2 e 3 concentram quase exclusivamente ocorrências sob tempo "bom", com porcentagens acima de 99%, enquanto o *Cluster* 0 apresenta maior diversidade, incluindo 34,2% de casos sob chuva e 50,4% em condições nubladas.

Na parte inferior esquerda, a matriz "Padrão Temporal: Período Detalhado" mostra a distribuição dos sinistros ao longo dos períodos do dia. Cada *cluster* exibe um padrão temporal marcante: o *Cluster* 1 é predominante no período matutino (99,1%), o *Cluster* 2 no período noturno (90,4%) e o *Cluster* 3 no vespertino (99,8%). O *Cluster* 4, por sua vez, apresenta concentração noturna mais equilibrada (53%), seguida por madrugada (15,5%).

A matriz inferior direita, "Padrão Temporal: Horário de Pico", destaca a proporção de sinistros ocorridos em horários de pico. Os *clusters* 1 e 2 concentram a cerca de 52% dos sinistros nesses horários, enquanto o *Cluster* 3 apresenta apenas 18% nesse contexto, indicando menor associação com congestionamentos.

Figura 39 - Distribuição dos Padrões Temporais por *Cluster*.



5.3.4. Aprendizagem de máquina supervisionado

Nesta seção, descreve-se os procedimentos realizados durante a execução do *script* de otimização balanceada para predição de variáveis relacionadas a sinistros de

trânsito. O objetivo principal foi treinar e validar modelos de *machine learning* supervisionado, tanto para tarefas de classificação quanto de regressão, garantindo a confiabilidade das predições através de validações cruzadas e alertas automáticos de *overfitting*.

O *dataset* foi dividido entre treino e teste utilizando a proporção de 75% para treino e 25% para teste, com estratificação nas tarefas de classificação.

Os seguintes algoritmos foram aplicados: *Random Forest Classifier, Extra Trees Classifier, Logistic Regression, XGBoost* e *LightGBM*. Para tarefas de regressão, foram aplicados: *Random Forest Regressor*, *Extra Trees Regressor*, *Linear Regression*, *Ridge Regression, XGBoost Regressor* e *LightGBM Regressor*.

Foram abordadas classes de problemas de classificação binária, multiclasse e regressão contínua. As variáveis-alvo incluem: período crítico do sinistro (manhã, tarde, noite, fim de semana), tipo de via (simples, dupla), gravidade e complexidade.

A avaliação dos modelos foi realizada por meio de validação cruzada (3-fold) para estimar a performance geral e evitar *overfitting*. Foram definidas métricas de alerta com base em limiares: *accuracy* > 0.98 e R^2 > 0.95 eram sinalizados como possivelmente inflacionados. Resultados com scores perfeitos (1.0) foram sinalizados como altamente suspeitos. Todos os modelos foram classificados com base na confiabilidade das predições.

A Tabela 9 consolida os principais resultados obtidos com os modelos preditivos considerados confiáveis, conforme critérios metodológicos definidos na etapa de validação cruzada. Destacam-se modelos com acurácia superior a 90%, como aqueles voltados à predição de condições da via (seca ou molhada) e de períodos compostos (ex.: manhã em fim de semana), todos validados com baixa variabilidade entre os *folds*.

Tabela 9 – Principais resultados obtidos.

Target	**Algoritmo**	**Tipo**	***Accuracy***	***F1-Score***	***CV Score***	***CV Std***
periodo_fds	*Logistic Regression*	Classificação	0.788	0.725	0.788	0.005
via_molhada	*Random Forest*	Classificação	0.963	0.957	0.963	0.004
via_seca	*Random Forest*	Classificação	0.959	0.96	0.959	0.004
periodo_fds_manhã_fds	*Logistic Regression*	Classificação	0.946	0.928	0.946	0.002
periodo_fds_manhã_semana	*Logistic Regression*	Classificação	0.936	0.943	0.936	0.002
periodo_fds_noite_fds	*Logistic Regression*	Classificação	0.932	0.907	0.932	0.001
periodo_fds_noite_semana	*Logistic Regression*	Classificação	0.927	0.938	0.927	0.004

periodo_fds_tarde_fds	*Logistic Regression*	Classificação	0.93	0.898	0.93	0.001
periodo_fds_tarde_semana	*LightGBM*	Classificação	0.93	0.931	0.924	0.01
via_dupla	*Random Forest*	Classificação	0.715	0.694	0.715	0.012
complexidade_acidente	*Random Forest*	Regressão	-	-	0.762	0.009

A aplicação de algoritmos de *clustering* em bases de dados complexas representa uma abordagem eficaz para identificar padrões latentes e estruturar perfis comportamentais em contextos em que as classes não são previamente definidas. No presente estudo, a técnica de *K-Means* foi utilizada para agrupar 4.601 registros de sinistros de trânsito em cinco *clusters* distintos, com base em um conjunto de 31 variáveis contínuas e categóricas previamente tratadas e codificadas. A redução de dimensionalidade por meio da Análise de Componentes Principais (PCA) foi aplicada para assegurar a eficiência do agrupamento, mantendo a maior variabilidade explicada com o menor número de componentes.

Uma vez definidos os *clusters*, não foram tratados como alvos das predições, o que configuraria um uso conceitualmente equivocado e metodologicamente inválido. Ao contrário, os *clusters* foram incorporados como variáveis explicativas (*features*) em modelos supervisionados, sendo utilizados como parte de uma estratégia avançada de engenharia de atributos (*feature engineering*).

Essa prática é documentada na literatura de ciência de dados e aprendizado de máquina, sendo recomendada pelo potencial de capturar efeitos combinados entre variáveis que não seriam detectados isoladamente, enriquecendo a capacidade preditiva dos modelos (KUHN; JOHNSON, 2019; JAMES *et al.*, 2021).

Três tarefas preditivas foram conduzidas a partir dessa estratégia: (i) a predição de perfis de risco associados aos sinistros, (ii) a estimativa de complexidade dos eventos e (iii) a classificação dos padrões temporais dos sinistros (manhã, tarde, noite ou madrugada). Em todos os casos, os *clusters* foram inseridos como uma variável categórica adicional no conjunto de preditores, ao lado de outras variáveis como horário, localização, tipo de veículo e condições climáticas.

Os resultados indicaram que a inclusão da variável *cluster* contribuiu para o aprimoramento dos modelos, com métricas robustas de desempenho: acurácia de 97,4% para classificação de risco com *Decision Tree*, coeficiente de determinação (R^2) de 0,867 na regressão de complexidade com *Random Forest*, e acurácia de 96,5% na classificação temporal com *Random Forest*. Essas evidências demonstram que os agrupamentos não

supervisionados forneceram insumos significativos para a predição supervisionada, atuando como uma forma de compressão semântica dos dados originais.

A utilização de *clusters* como atributos preditores, e não como alvos das predições, configura um recurso que amplia a capacidade explicativa dos modelos, favorecendo a construção de sistemas preditivos mais precisos, interpretáveis e aplicáveis em contextos reais de análise como no caso da gestão da segurança viária.

Análise Crítica das Predições com alto score

O uso de algoritmos de aprendizado de máquina em tarefas supervisionadas, como classificação e regressão, visa construir modelos capazes de generalizar padrões e realizar previsões com boa acurácia sobre dados novos. No entanto, quando os modelos apresentam escores muito elevados, como *accuracy* ou coeficiente de determinação (R^2) próximos ou iguais a 1.000, é necessário adotar uma abordagem criteriosa para distinguir entre bom desempenho real e *overfitting*, isto é, a memorização dos dados de treino, o que compromete a capacidade de generalização.

O sistema de otimização balanceada utilizado neste projeto destaca-se justamente por implementar esse cuidado. Ele mantém a análise original, mas adiciona validações e alertas inteligentes para interpretar corretamente o significado desses altos scores. A partir dessa lógica, os resultados foram classificados como:

- Confiáveis: resultados altos e realistas, com variações esperadas e boa performance em validação cruzada.
- Suspeitos: resultados possivelmente inflacionados, com *scores* perfeitos ou muito próximos de 1.0, que levantam suspeita de *overfitting* ou problemas como *data leakage*, classes desbalanceadas ou codificações excessivamente informativas.

Predições confiáveis são aquelas cujos modelos apresentaram desempenho elevado e consistente entre os dados de treino e teste (validação cruzada), com métricas realistas. No presente estudo, destacam-se 11 predições consideradas confiáveis seguindo estes parâmetros (Tabela 10). Esses resultados indicam que, mesmo com valores elevados, os scores não são perfeitos, o que é um sinal saudável de que o modelo está lidando com variação nos dados e não está decorando padrões específicos.

Tabela 10 - 11 modelos classificados como confiáveis.

N	Atributo	Accuracy (ou R^2)	Algoritmo
1	periodo_fds	0,788	*Logistic Regression*
2	via_molhada	0,963	*Random Forest*
3	via_seca	0,959	*Random Forest*
4	periodo_fds_manhã_fds	0,946	*Logistic Regression*
5	periodo_fds_manhã_semana	0,936	*Logistic Regression*
6	periodo_fds_noite_fds	0,932	*Logistic Regression*
7	periodo_fds_noite_semana	0,927	*Logistic Regression*
8	periodo_fds_tarde_fds	0,93	*Logistic Regression*
9	periodo_fds_tarde_semana	0,924	*LightGBM*
10	via_dupla	0,715	*Random Forest*
11	complexidade_acidente	0,762	*Random Forest*

Na prática, as predições geradas por modelos de *machine learning* em problemas como o analisado neste estudo podem servir como instrumentos poderosos de apoio à decisão, sobretudo em contextos de gestão pública, segurança viária, planejamento urbano e prevenção de riscos.

5.4. Conclusão

A aplicação do sistema de detecção e supressão de vazamento de dados (*data leakage*) resultou na exclusão de 27 variáveis temporais e identificadores que permitiam inferências indevidas, preservando 18 preditores livres de viés temporal. Essa escolha metodológica assegurou que os modelos gerados refletissem relações genuínas entre condições da via, fatores sazonais e infraestrutura, em vez de memorizarem padrões específicos de séries históricas.

Com a base depurada, foram desenvolvidos onze modelos supervisionados considerados confiáveis, cujas métricas de desempenho oscilaram entre 67,8% e 96,7% de acurácia ou R^2, aferidas por validação cruzada estratificada. Em paralelo, a análise não supervisionada identificou cinco *clusters* distintos de sinistros, complementando a compreensão dos padrões de risco. A diversidade de alvos, como condições adversas da via, gravidade detalhada e presença de pista dupla ou molhada, demonstra capacidade de resposta a distintos quesitos operacionais de segurança viária.

A análise de importância das variáveis indicou a centralidade de fatores geográficos, sazonais e de infraestrutura na ocorrência dos sinistros, reforçando o predomínio de padrões espaciais e ambientais sobre marcadores exclusivamente

temporais. Ainda assim, a remoção de atributos temporais de alta cardinalidade impôs limitação intrínseca: a impossibilidade de prever janelas horárias específicas ou variações semanais muito finas, um trade-off metodológico necessário para manter a validade preditiva em cenários prospectivos.

Por fim, o trabalho evidencia que procedimentos rigorosos de saneamento de dados e controle de *data leakage* constituem pré-requisito para modelos de aprendizado de máquina que se pretendem generalizáveis e úteis na gestão de segurança viária. Os resultados estabelecem um referencial metodológico que pode ser transposto a outros contextos em que a distinção entre correlação legítima e vazamento de informação é crucial para a credibilidade científica e a aplicação prática dos modelos.

5.5. Referências Bibliográficas

JORDAN, Michael I.; MITCHELL, Tom M. **Machine learning: trends, perspectives, and prospects**. Science, Washington, v. 349, n. 6245, p. 255-260, 2015.

KAUFMAN, S.; ROSSET, S.; PERLICH, C.; STITELMAN, O. **Leakage in data mining: formulation, detection, and avoidance**. ACM Transactions on Knowledge Discovery from Data, New York, v. 6, n. 4, p. 1-21, 2012. DOI: 10.1145/2382577.2382579.

KAPOOR, S.; NARAYANAN, A. **Leakage and the reproducibility crisis in machine-learning-based science**. Patterns, [S. l.], v. 4, n. 9, p. 100804, 2023. DOI: 10.1016/j.patter.2023.100804.

MCCARTT, Anne T.; MAYHEW, Dan R.; BRAITMAN, Keli A.; FERGUSON, Susan A.; SIMPSON, Herb M. **Effects of age and experience on young driver crashes**: review of recent literature. Traffic Injury Prevention, v. 10, n. 3, p. 209–219, 2009. https://doi.org/10.1080/15389580802677807.

FERNÁNDEZ, Alberto; GARCÍA, Salvador; GALAR, Mikel; PRATI, Ronaldo C.; KRAWCZYK, Bartosz; HERRERA, Francisco. **Learning from imbalanced data sets**. Cham: Springer, 2018. https://doi.org/10.1007/978-3-319-98074-4.

MICCI-BARRECA, Dan. **A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems**. ACM SIGKDD Explorations Newsletter, v. 3, n. 1, p. 27–32, 2001. https://doi.org/10.1145/507533.507538.

NEWMAN, M. E. J. **Power laws, Pareto distributions and Zipf's law**. Contemporary Physics, v. 46, n. 5, p. 323–351, 2005. https://doi.org/10.1080/00107510500052444.

BISHOP, Christopher M. **Pattern recognition and machine learning**. New York: Springer, 2006.

MACQUEEN, James. **Some methods for classification and analysis of multivariate observations**. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967, Berkeley. Proceedings. Berkeley: University of California Press, 1967. p. 281-297.

ESTER, Martin et al. **A density-based algorithm for discovering clusters in large spatial databases with noise**. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2., 1996, Portland. Proceedings… Menlo Park: AAAI Press, 1996. p. 226-231.

WARD, Joe H. **Hierarchical grouping to optimize an objective function**. Journal of the American Statistical Association, Alexandria, v. 58, n. 301, p. 236-244, 1963.

JOLLIFFE, Ian T. **Principal component analysis**. 2. ed. New York: Springer, 2002.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. **Isolation forest**. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 8., 2008, Pisa. Proceedings. Los Alamitos: IEEE, 2008. p. 413-422.

BREIMAN, Leo. **Random forests**. Machine Learning, Dordrecht, v. 45, n. 1, p. 5-32, 2001.

GEURTS, Pierre; ERNST, Damien; WEHENKEL, Louis. **Extremely randomized trees.** Machine Learning, Dordrecht, v. 63, n. 1, p. 3-42, 2006.

HOSMER, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied logistic regression**. 3. ed. Hoboken: Wiley, 2013.

MURPHY, Kevin P. **Machine learning: a probabilistic perspective**. Cambridge: MIT Press, 2012.

CHEN, Tianqi; GUESTRIN, Carlos. **XGBoost: A scalable tree boosting system**. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. Proceedings. New York: ACM, 2016. p. 785-794.

KE, Guolin et al. LightGBM: **A highly efficient gradient boosting decision tree**. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017, Long Beach. Proceedings. [S.l.: s.n.], 2017. p. 3149-3157.

KUHN, Max; JOHNSON, Kjell. **Feature engineering and selection: a practical approach for predictive models**. Boca Raton: CRC Press, 2019.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An introduction to statistical learning: with applications in R**. 2. ed. New York: Springer, 2021.

6. CONCLUSÃO GERAL

A presente dissertação investiga, de modo integrado, os sinistros de trânsito ocorridos no município de Rio Verde (GO) entre 1.º de janeiro de 2021 e 31 de dezembro de 2024, período em que a Agência Municipal de Mobilidade e Trânsito (AMT) registrou 7.926 ocorrências eletrônicas. Tal recorte temporal coincide com o primeiro quadriênio da Década de Ação pela Segurança no Trânsito 2021-2030, bem como com a fase inicial da consolidação do Plano Nacional de Redução de Mortes e Lesões no Trânsito (PNATRANS), ambos referenciais normativos que orientam políticas públicas de prevenção no Brasil e em escala global.

À luz desses marcos, o trabalho foi estruturado em três artigos complementares que, em conjunto, respondem ao objetivo geral de caracterizar padrões estatísticos, espaciais e preditivos dos sinistros locais. O primeiro artigo desenvolveu análise exploratória e estatística descritiva da série temporal; o segundo examinou a distribuição espacial e espaciotemporal por técnicas de densidade *kernel*, *Moran* I/LISA e DBSCAN; o terceiro avaliou a capacidade de modelos de aprendizado de máquina, com controle rigoroso de *data leakage*, antever condições de risco com base em variáveis operacionais.

Os resultados consolidados demonstram que a integração de análise exploratória, estatística espacial e modelagem preditiva, aplicada a uma base eletrônica municipal com alta completude, constitui abordagem suficiente para caracterizar padrões, identificar áreas críticas e antecipar condições de risco de sinistros de trânsito.

Ao mesmo tempo, reforça a aderência do município às metas do PNATRANS e à Década de Ação pela Segurança no Trânsito 2021-2030, fornecendo um modelo replicável de governança de dados e monitoramento contínuo. Dessa forma, a dissertação cumpre o objetivo de gerar conhecimento aplicável e delineia um caminho claro para a evolução de políticas públicas baseadas em evidências, sustentado por metodologias transparentes e reprodutíveis.

Limitações do estudo

Apesar da elevada completude da base eletrônica e da abrangência temporal de quatro anos, o estudo apresenta limitações que condicionam a interpretação dos resultados. Persistem indícios de subnotificação, já que incidentes sem acionamento oficial, sobretudo aqueles com danos leves e ausência de vítimas, não integram o registro

da AMT, gerando possível viés de representatividade. A precisão geoespacial, embora alta, é heterogênea: parte das ocorrências foi geocodificada na sede da AMT, e resultou em limpeza destes registros, a cerca de 10% do total do *dataset,* resultando na dificuldade de análises geoespaciais por bairro.

Os modelos preditivos, ainda que controlados para *data leakage*, foram limitados a dezoito preditores operacionais, uma vez que o processo de detecção e supressão excluiu variáveis temporais e identificadores que permitiam inferências indevidas, reduzindo o conjunto original de 44 atributos para 18 efetivamente utilizáveis, o que impôs um teto ao desempenho alcançado ao excluir fatores comportamentais e infraestruturais mais finos.

Ademais, o recorte territorial restrito a Rio Verde e as particularidades de sua malha viária e fiscalização limitam a generalização direta dos achados para municípios com características distintas. Por fim, o horizonte temporal de 2021 a 2024 permite capturar variações sazonais, mas pode não evidenciar mudanças estruturais de longo prazo decorrentes de intervenções graduais de engenharia ou de alterações macroeconômicas que influenciem o volume de tráfego.

Benefícios operacionais preliminares para a Agência Municipal de Mobilidade e Trânsito (AMT)

A execução deste projeto já resulta em ganhos tangíveis para a AMT. Historicamente, os quase oito mil registros de sinistros referentes ao período de 2021–2024 eram inseridos manualmente por servidores da autarquia, procedimento moroso e propenso a erros. A presente pesquisa desenvolveu e destacou, no Capítulo I, um *script* de extração automática dos campos constantes nos formulários eletrônicos; esse código será entregue à AMT em formato reproduzível, reduzindo o tempo de processamento de novos registros e padronizando a qualidade da base.

Além da automação de coleta, serão fornecidos mapas interativos dos sinistros em ambiente web, que podem ser integrados aos painéis de BI já utilizados pelo órgão, suprindo uma lacuna na análise espacial atualmente inexistente nos BI internos e oferecendo suporte imediato à tomada de decisão.

A análise das coordenadas revelou, ainda, oportunidade de aprimoramento na etapa de georreferenciamento: mais de dez por cento dos registros apresentam localização coincidente com a sede da AMT. Tal padrão sugere duas hipóteses complementares: preenchimento integral dos boletins nas dependências da autarquia ou finalização dos

formulários apenas após retorno à base, e aponta para a necessidade de reforçar a captura de coordenadas no local do sinistro. A adoção do *script* e dos mapas, aliados a orientações operacionais, tende a mitigar esse viés geográfico, elevando a confiabilidade da informação para fins de planejamento e fiscalização.

REFERÊNCIAS BIBLIOGRÁFICAS

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). **Plano Global para a Década de Ação pela Segurança no Trânsito 2021-2030**. Genebra: Organização Mundial da Saúde, 2020.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Relatório mundial sobre a situação da segurança no trânsito 2019**. Genebra: Organização Mundial da Saúde, 2019.

BRASIL. Ministério da Saúde. Departamento de Informática do SUS – DATASUS. **Mortalidade por sinistros de transporte terrestre, Goiás, 2020-2021**. Brasília, 2023. Disponível em: http://www2.datasus.gov.br. Acesso em: 03 jun. 2025.

SERVIÇO FEDERAL DE PROCESSAMENTO DE DADOS (SERPRO). Sistema RENAVAM. **Relatório Estatístico de Sinistros de Trânsito – RENAEST**: Rio Verde-GO, 2021-2024. Brasília: SENATRAN, 2024. Disponível em: https://www.gov.br/transportes/pt-br/assuntos/transito/arquivos-senatran/docs/renaest. Acesso em: 03 jun. 2025.

AMOROS, E.; MARTIN, J.-L.; LAUMON, B. **Under-reporting of road crash casualties in France**. Accident Analysis & Prevention, v. 38, n. 4, p. 627–635, 2006.

ALSOP, J.; LANGLEY, J. **Under-reporting of motor vehicle traffic crash victims in New Zealand. Accident Analysis & Prevention**, v. 33, n. 3, p. 353-359, 2001.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Global status report on road safety 2021**. Genebra: Organização Mundial da Saúde, 2021.

CHANG, L.-Y.; CHEN, W.-C. **Data mining of tree-based models to analyze freeway accident frequency**. Journal of Safety Research, v. 36, n. 4, p. 365-375, 2005.

LORD, D.; MANNERING, F. **The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives**. Transportation Research Part A: Policy and Practice, v. 44, n. 5, p. 291-305, 2010.

KAUFMAN, S. et al. **Leakage in data mining: formulation, detection, and avoidance**. ACM Transactions on Knowledge Discovery from Data, v. 6, n. 4, p. 1-21, 2012.

KAPOOR, S.; NARAYANAN, A. **Leakage and the reproducibility crisis in machine-learning-based science**. Patterns, v. 4, n. 9, p. 100804, 2023.

BRASIL. Ministério dos Transportes. **Plano Nacional de Redução de Mortes e Lesões no Trânsito (PNATRANS)**. Brasília, 2023. Disponível em: https://www.gov.br/transportes/pt-br/assuntos/transito/pnatrans. Acesso em: 16 jun. 2025.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). **Plano Global para a Década de Ação pela Segurança no Trânsito 2021-2030**. Genebra: Organização Mundial da Saúde, 2020.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). **Transformando nosso mundo: a Agenda 2030 para o Desenvolvimento Sustentável**. Resolução A/RES/70/1, Nova Iorque, 2015.

WASHINGTON, Sean P.; KARLAFTIS, Matthew G.; MANNERING, Fred L. **Statistical and econometric methods for transportation data analysis**. 3. ed. Boca Raton: CRC Press, 2020.

ELVIK, Rune; HØYE, Alena; VAA, Truls; SØRENSEN, Michael. **The handbook of road safety measures**. 2. ed. Bingley: Emerald, 2009.

MONTGOMERY, Douglas C.; RUNGER, George C. **Applied statistics and probability for engineers**. 7. ed. Hoboken: John Wiley & Sons, 2018.

O'SULLIVAN, D.; UNWIN, D. **Geographic information analysis**. Hoboken: Wiley, 2003.

SILVERMAN, B. W. **Density estimation for statistics and data analysis**. London: Chapman and Hall, 1986.

XIE, Z.; YAN, J. **Detecting traffic accident clusters with network kernel density estimation and local spatial autocorrelation**. Accident Analysis and Prevention, v. 50, p. 477-486, 2013.

ANSELIN, L. **Local indicators of spatial association** – LISA. Geographical Analysis, v. 27, n. 2, p. 93-115, 1995.

ESTER, M. et al. **A density-based algorithm for discovering clusters in large spatial databases with noise**. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996.

CHAINEY, S.; RATCLIFFE, J. **GIS and crime mapping**. Chichester: Wiley, 2013.

RUSSEL, S., NORVIG, P. **Artificial Intelligence: A Modern Approach**. New Jersey: Pearson Education, 2013.

GOMES, DOS SANTOS. **Inteligência Artificial: Conceitos e Aplicações**. Revista Olhar Científico – Faculdades Associadas de Ariquemes – v. 01, p. 234-246, 2010.

MITCHELL, T. **Machine Learning**. McGraw-Hill, 1997.

IZBICKI, R., DOS SANTOS. **Aprendizado de Máquina: Uma Abordagem Estatística**. São Carlos, SP: Rafael Izbicki, 2022.

MÜLLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. Sebastopol: O'Reilly Media, 2016.

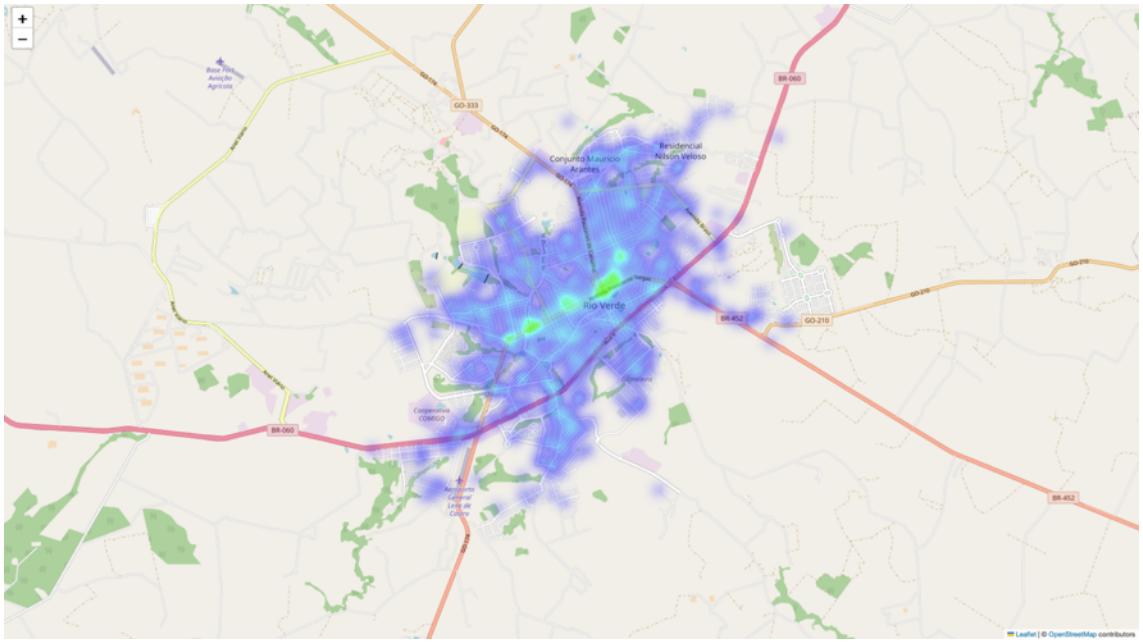
APÊNDICES

Apêndice A

Mapa interativo de calor – Sinistros em Rio Verde (GO) - 2021 a 2024

Link: https://drive.google.com/file/d/1skBg4roT6lB3RGUbNf103mYwYa7NEmLN/view?usp=share_link

Orientação: Para visualizar o conteúdo, é necessário baixar o arquivo em formato HTML e abri-lo em um navegador.

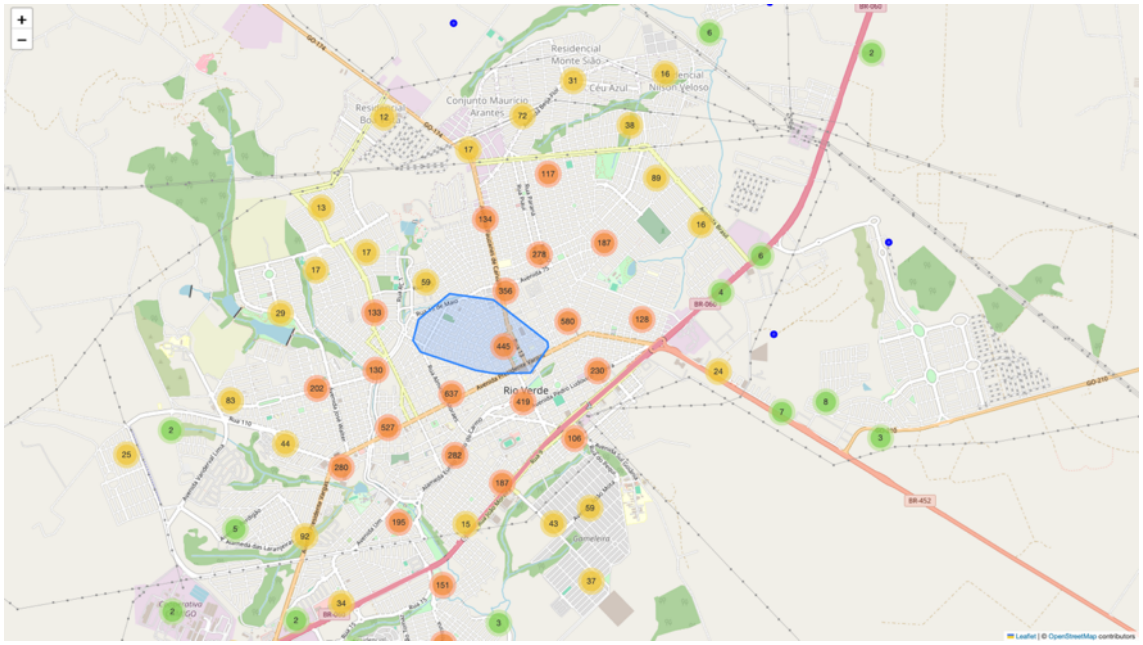


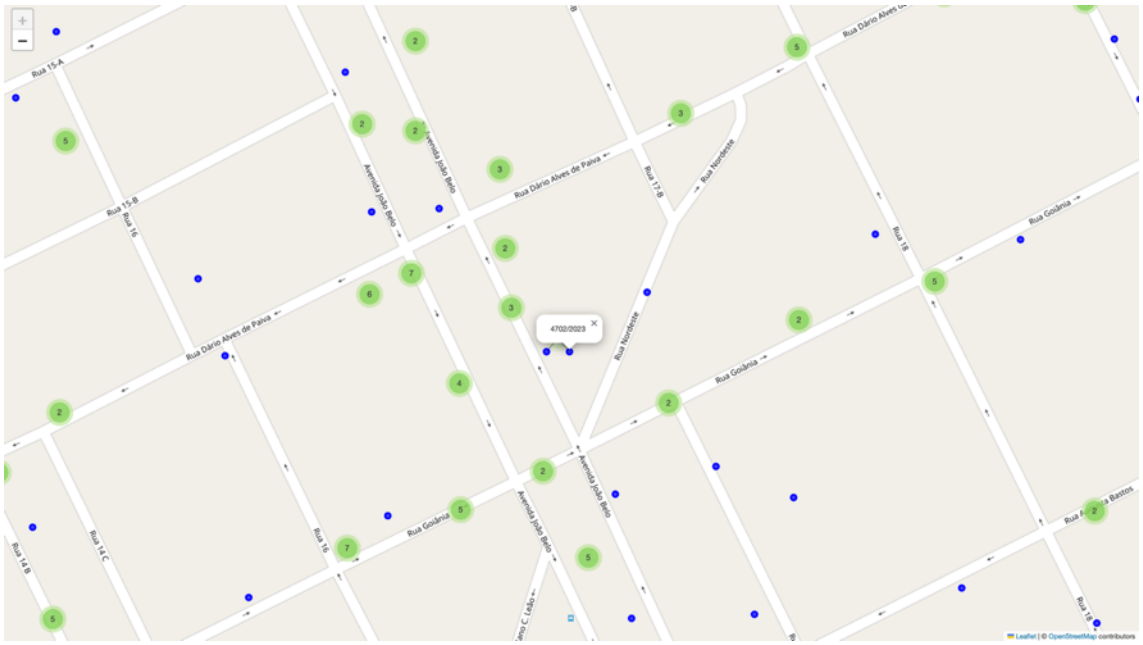
Apêndice B

Mapa interativo por agrupamento – Sinistros em Rio Verde (GO) - 2021 a 2024 por número do Boletim

Link: https://drive.google.com/file/d/1LI28TfSZcrf2tRvQNopIgJnO9ngvwE1E/view?usp=share_link

Orientação: Para visualizar o conteúdo, é necessário baixar o arquivo em formato HTML e abri-lo em um navegador.





Apêndice C

Mapa interativo de calor – Sinistros na Avenida Presidente Vargas, Rio Verde (GO) 2021 a 2024

Link: https://drive.google.com/file/d/1gZe2lKSfcQwztg9xpbAag9xDb4ulAx-U/view?usp=share_link

Orientação: Para visualizar o conteúdo, é necessário baixar o arquivo em formato HTML e abri-lo em um navegador.