



**Universidade Federal de Santa Catarina
Centro Tecnológico
Departamento de Engenharia Civil**

Diego Wainstein Silva

**Estabelecimento de relações entre postos de monitoramento de tráfego com base em
dados geográficos e algoritmos de clusterização**

11 de julho de 2025

Diego Wainstein Silva

Estabelecimento de relações entre postos de monitoramento de tráfego com base em dados geográficos e algoritmos de clusterização

Trabalho de Conclusão de Curso apresentado pelo acadêmico Diego Wainstein Silva à banca examinadora do Curso de Graduação em Engenharia Civil da Universidade Federal de Santa Catarina como requisito parcial para obtenção do título de Engenheiro Civil.

Professor orientador: Alexandre Hering Coelho, Dr.

11 de julho de 2025

Diego Wainstein Silva

Estabelecimento de relações entre postos de monitoramento de tráfego com base em dados geográficos e algoritmos de clusterização

Este Trabalho de Conclusão de Curso foi julgado adequado para a obtenção do Título de Engenheira Civil e aprovado em sua forma final pelo Departamento de Engenharia Civil da Universidade Federal de Santa Catarina

11 de julho de 2025

Prof. Alexandre Hering Coelho, Dr.
Coordenador do Curso

Banca examinadora:



Documento assinado digitalmente

Alexandre Hering Coelho

Data: 11/07/2025 17:04:34-0300

CPF: ***.478.099-**

Verifique as assinaturas em <https://v.ufsc.br>

Prof. Alexandre Hering Coelho, Dr.
Orientador

Prof. Liseane Padilha Thives, Dra.
Membro

Prof. Luciana Rohde, Dra.
Membro

Agradecimentos

Agradeço ao meu orientador Professor Doutor Alexandre Hering Coelho pelos diversos ensinamentos durante o desenvolvimento do TCC e nas disciplinas que me lecionou ao longo do curso.

Meus pais, avós, tia e minha irmã que me deram exemplo e suporte para estudar.

Minha namorada por me dar apoio e incentivo quando foi necessário.

Aos demais familiares e aos meus amigos que me ajudaram através de companheirismo e conselhos.

Obrigado.

RESUMO

O monitoramento do tráfego rodoviário desempenha papel fundamental no planejamento e manutenção da infraestrutura viária. Diante da dificuldade de realizar contagens contínuas em todos os pontos da malha, sobretudo por razões econômicas, surge a necessidade de estabelecer relações entre postos de contagem permanentes e postos de cobertura, de modo a permitir a expansão volumétrica. Este trabalho tem como objetivo agrupar postos de monitoramento de tráfego utilizando o algoritmo de clusterização *K-means*, com base em variáveis geográficas, e, a partir dos grupos formados, associar a cada posto de cobertura o posto permanente mais semelhante. Para isso, foi adotado um método híbrido, conforme a classificação da FHWA (2022), que combina uma abordagem *knowledge-based* para seleção das variáveis com análise de agrupamento estatístico. No processo de construção das variáveis, é aplicado o conceito de áreas de influência, definidas por meio de *buffers* espaciais em torno dos postos, dentro dos quais foram calculadas métricas geográficas visando à caracterização da região do posto. A associação dos postos de cobertura aos de coleta contínua é realizada por similaridade euclidiana no espaço das variáveis normalizadas. Os resultados mostraram que o agrupamento apresentou coerência espacial e interpretativa, sendo possível estabelecer um perfil para cada *cluster*. Foram formados quatro grupos: o primeiro representa áreas economicamente relevantes com tráfego moderado e infraestrutura consolidada; o segundo, regiões menos dinâmicas e com infraestrutura básica; o terceiro reúne áreas urbanas densas e de tráfego intenso; e o quarto corresponde a trechos planejados, em contextos estaduais, com menor maturidade de infraestrutura. Variáveis relacionadas à infraestrutura e à jurisdição foram as mais relevantes estatisticamente na formação dos grupos. É concluído que os dados geográficos são suficientes para agrupar postos com perfis marcados, e que o método adotado se mostrou replicável com ferramentas acessíveis. Como limitações, destaca-se a não utilização de dados temporais de volume de tráfego e a aplicação de um único algoritmo de agrupamento. Estudos futuros podem incorporar algoritmos alternativos, análise temporal por DTW, validação estatística dos grupos e técnicas como PCA para aprimorar a robustez e a aplicabilidade do método proposto. Assim, este trabalho estabelece uma base promissora, mas que demanda aprofundamento técnico para ser plenamente aproveitada como ferramenta de apoio à expansão volumétrica e à gestão inteligente de infraestrutura viária.

Palavras-chave: engenharia de tráfego, postos de monitoramento de tráfego, expansão volumétrica, *clusters*, *K-means*

ABSTRACT

Road traffic monitoring plays a fundamental role in the planning and maintenance of transportation infrastructure. Given the difficulty of performing continuous counts at all points of the network, especially due to economic constraints, there arises a need to establish relationships between permanent counting stations and coverage stations, enabling volumetric expansion. This study aims to cluster traffic monitoring stations using the K-means algorithm, based on geographic variables, and, from the formed groups, associate each coverage station with the most similar permanent station. To this end, a hybrid method was adopted, following the classification of FHWA (2022), which combines a knowledge-based approach for variable selection with statistical clustering analysis. In the variable construction process, the concept of areas of influence was applied, defined through spatial buffers around the stations, within which geographic metrics were calculated to characterize the station's region. The association of coverage stations with continuous counting stations was performed using Euclidean similarity in the space of normalized variables. The results showed that the clustering presented spatial and interpretative coherence, allowing the definition of a profile for each cluster. Four groups were formed: the first represents economically relevant areas with moderate traffic and consolidated infrastructure; the second includes less dynamic regions with basic infrastructure; the third comprises dense urban areas with intense traffic; and the fourth corresponds to planned sections, in state-level contexts, with lower infrastructure maturity. Variables related to infrastructure and jurisdiction were statistically the most relevant in forming the groups. It is concluded that geographical data are sufficient to cluster stations with distinct profiles, and that the adopted method proved to be replicable using accessible tools. As limitations, the lack of temporal traffic volume data and the application of only a single clustering algorithm are highlighted. Future studies may incorporate alternative algorithms, temporal analysis using DTW, statistical validation of the groups, and techniques such as PCA to enhance the robustness and applicability of the proposed method. Thus, this study establishes a promising foundation that requires further technical development to be fully leveraged as a tool to support volumetric expansion and the intelligent management of road infrastructure.

Key-words: traffic engineering, traffic monitoring stations, traffic volume expansion, clusters, K-means

Lista de Figuras

1	Relação entre a hora e o volume horário de tráfego em rodovias norte-americanas (DNIT/IPR, 2006)	5
2	Equipamento contador em postos de coleta contínua do PNCT (DNIT, 2018b)	8
3	Explicação do equipamento contador em postos de coleta contínua do PNCT (DNIT, 2018b)	8
4	Variações horárias típicas em vias americanas (DNIT/IPR, 2006)	11
5	Variação mensal em rodovias rurais e urbanas (DNIT/IPR, 2006)	12
6	Aplicação de fatores de expansão para transformar VMD horário, semanal e mensal (sazonal) para VMD_a (DNIT, 2008)	13
7	Método do cotovelo em conjunto exemplo de dados (ALMEIDA; CARVALHO; MENINO, 2017)	24
8	Método da silhueta (ALMEIDA; CARVALHO; MENINO, 2017)	25
9	Posicionamento dos centroides e atribuição dos pontos (MÜLLER; GUIDO, 2016) . .	26
10	Amostra da tabela de postos	32
11	Postos de monitoramento de tráfego e rodovias	33
12	Cotovelo	35
13	Índice de silhueta	36
14	Importância das variáveis no agrupamento	37
15	Dispersão dos postos por <i>clusters</i> - eixos: população $\times$ acidentes	38
16	Dispersão dos postos e centroides por <i>clusters</i> - eixos: população $\times$ acidentes . . .	39
17	Distribuição espacial dos postos, coloridos por <i>cluster</i>	40

Lista de Tabelas

1	Atributos categóricos de um posto exemplo	33
2	Codificação binária dos atributos categóricos	33
3	Quantidade de postos pertencentes a cada <i>cluster</i>	37
4	Médias das variáveis para cada <i>cluster</i> de postos	39
5	Postos de cobertura e o posto permanente mais similiar	42

Sumário

1	Introdução	1
1.1	Objetivos	2
1.1.1	Objetivos gerais	2
1.1.2	Objetivos específicos	2
2	Revisão bibliográfica	3
2.1	Caracterização da demanda de tráfego em rodovias rurais	3
2.1.1	Contagem de tráfego	6
2.1.2	Expansão volumétrica	9
2.2	Estabelecimento de relações entre postos	14
2.3	Aprendizado de máquinas	17
2.4	Agrupamento	19
2.4.1	<i>K-means</i>	22
3	Método	27
4	Aplicação	28
4.1	Aquisição e preparação de dados	28
4.2	Agrupamento dos postos	34
5	Resultados	35
6	Conclusões	43
6.1	Sugestões de trabalhos futuros	43
	Referências	45
	Anexos	47
A	Códigos de <i>Python</i>	47
A.1	Aquisição de dados	47
A.2	Preparação de dados	51
A.3	Agrupamento	52
A.4	Cálculos sobre os grupos	55

1 Introdução

A Confederação Nacional do Transporte (CNT) afirma que o modo de transporte rodoviário representa 65% da movimentação de carga e 95% dos passageiros no Brasil (CNT; SEST SENAT, 2023). A mais recente pesquisa de CNT e SEST SENAT (2024) avaliou 111.853 km de rodovias e, através da análise de pavimentação, geometria e sinalização, classificou o estado geral da malha como 67% regular, ruim ou péssimo, e apenas 33% como bom ou ótimo.

Essa precariedade estrutural ocorre em paralelo ao crescimento da demanda: de 2013 a 2023, a frota nacional de veículos aumentou 46,1%, passando de 81,6 para 119,2 milhões de unidades (Ministério dos Transportes, 2023). Diante desse cenário, torna-se ainda mais urgente o planejamento de melhorias, reformas e ampliação da infraestrutura viária, o que requer decisões baseadas em dados confiáveis e sistemáticos sobre o tráfego.

Para suprir essa demanda, o Departamento Nacional de Estradas de Rodagem (DNER) criou, em 1975, o “Programa de Contagem Sistemática de Trânsito”, que evoluiu até abranger 285 postos de monitoramento espalhados pelo país. Após sua interrupção em 2001, o Departamento Nacional de Infraestrutura de Transportes (DNIT) retomou a iniciativa em 2013, dando continuidade ao Programa Nacional de Contagem de Tráfego (PNCT), que atualmente conta com 365 postos de contagem contínua e 65 de cobertura. Esses dois tipos de posto são detalhados na Seção 2.1.1.

Historicamente, a classificação dos postos sempre foi realizada por critérios fixos e conhecimento prévio (abordagem *knowledge-based*) (SCHNEIDER; TSAPAKIS, 2009). Um exemplo é o trabalho de Bellamy (1978), que agrupou os postos em quatro categorias principais: rodovias recreacionais, não recreacionais de baixo fluxo, rurais de longa distância e urbanas de trajeto diário (ou pendular). No entanto, essas classificações careciam de flexibilidade e atualização frente às transformações geográficas e socioeconômicas do território brasileiro.

Com os avanços computacionais, de *machine learning* e a maior disponibilidade de dados georreferenciados, tornaram-se viáveis métodos baseados em dados e algoritmos de agrupamento. Esses métodos permitem identificar grupos de postos com características geográficas semelhantes a partir de múltiplas variáveis, sem a necessidade de categorias pré-definidas. A clusterização, portanto, emerge como uma ferramenta útil tanto para a compreensão dos padrões espaciais do PNCT quanto para o apoio à expansão volumétrica por meio da criação de grupos de fatores de ajuste sazonais (SAFs), conforme discutido por FHWA (2022).

Dessa forma, este trabalho investiga o uso de algoritmos de agrupamento para melhorar a análise e a gestão do PNCT, com foco na associação entre postos de monitoramento de tráfego a partir de suas características geográficas.

1.1 Objetivos

Nesta seção são apresentados os objetivos do trabalho.

1.1.1 Objetivos gerais

O objetivo geral do trabalho é agrupar postos de monitoramento de tráfego por meio do algoritmo *K-means* e associar, a cada posto de cobertura, o posto de coleta contínua mais similar, utilizando dados geográficos.

1.1.2 Objetivos específicos

Os objetivos específicos são listados a seguir.

- a) descrever um perfil para cada grupo de postos do PNCT formado;
- b) avaliar a dificuldade de encontrar dados sistemáticos e abrangentes;
- c) analisar o grau de dificuldade exigido para o processamento dos dados;
- d) determinar as variáveis mais importantes na formação dos grupos;
- e) verificar o quão acessíveis são os recursos de software disponíveis para auxiliar em tarefas de clusterização atualmente;
- f) avaliar quais são os principais benefícios de agrupar postos de monitoramento de tráfego utilizando somente variáveis geográficas.

2 Revisão bibliográfica

Este capítulo visa trazer uma fundamentação teórica sobre a demanda de tráfego, sua contabilização através de postos de monitoramento e a relevância técnica do agrupamento dos postos. Introduz o conceito de *machine learning* e suas principais técnicas e aplicações e traz uma lista dos tipos de algoritmos de agrupamento, é explicado, mais detalhadamente, o algoritmo *K-means*, cobrindo deste, aquilo que é usado no método do presente trabalho.

2.1 Caracterização da demanda de tráfego em rodovias rurais

O processo de caracterização da demanda de tráfego, de rodovias em geral, envolve a identificação e análise de diversos fatores que influenciam a composição do tráfego (tipos de veículo), a origem e o destino das viagens, o volume e suas variações com o tempo (explicado na Seção 2.1.2).

- **Composição do tráfego:** uma frota de veículos difere entre si quanto a tamanho, peso e velocidade. Algumas classificações, como a feita para projetos, segundo o manual de estudos de tráfego (DNIT/IPR, 2006) leva em consideração largura, comprimento e o raio das rodas para diferenciar os veículos, já a classificação do DNIT utiliza o número de eixos. Esta classificação é usada, por exemplo, para estipular um coeficiente multiplicador da tarifa de pedágios. Existe também classificação para cálculo de capacidade, visto que duas rodovias de geometria igual podem ter capacidades diferentes, isso porque o tipo de veículos que transita por elas é diferente. O conhecimento mais relevante para tirar disso é que a composição do tráfego pode ser medida de diversas formas e para cada contexto terá um modo de fazê-la associado ao seu objetivo.
- **Origem/Destino:** baseia-se em dividir uma área de estudo em zonas de tráfego e contar, para cada viagem feita dentro da área de estudo, qual foi a zona de origem e de destino dela. Dessa forma obtendo zonas de tráfego de 'geração' e de 'atração' de viagens. Normalmente, zonas residenciais são de geração, ou seja, são ponto de origem de muitas viagens, e áreas turísticas ou com alta concentração de ambientes de trabalho são zonas de atração, que são o destino das viagens, por exemplo. Determinar o grau de geração e atração de viagens é um número necessário, segundo DNIT/IPR (2006), para projetar o tráfego futuro de uma região, o que fornece elementos para dimensionamento dos projetos de engenharia, visto que este é feito para atender um nível de serviço para toda a vida útil da infraestrutura, ou seja, anos no futuro.
- **Volume:** é número de veículos contados no ponto de observação dividido pelo tempo total observado e é medido em veículos por unidade de tempo, por exemplo, veículos por hora (v/h).

Da contabilização do volume surgem os seguintes valores usados na engenharia (DNIT/IPR, 2006).

- Volume médio diário anual (VMD_a) - quantidade de veículos observados trafegando ao longo de um ano dividido por 365 (ou 366) dias do ano.
- Volume médio diário mensal (VMD_m) - quantidade de veículos observados trafegando ao longo de um mês dividido por 30 (ou 28, 29 ou 31) dias do mês. Deve-se indicar a qual mês os valores correspondem.
- Volume médio diário semanal (VMD_s) - quantidade de veículos observados trafegando ao longo de uma semana dividido por 7 dias da semana. Deve-se indicar a qual mês os valores correspondem.
- Volume médio diário em um dia de semana (VMD_d) - quantidade de veículos observados trafegando em um dia da semana. Deve-se indicar a qual dia da semana e mês os valores correspondente (ex.: terças-feiras de maio).
- Volume horário (VH) - quantidade de veículos observados trafegando em uma hora (existe também a medida para meia hora e para quarto de hora em alguns casos).

Do VH é derivado o volume horário de projeto (VHP). São ordenadas todas as horas do ano pelo volume de veículos trafegando na rodovia e, a depender do custo benefício visado, estipula-se o volume horário de projeto, que é aquele que será usado para dimensionamento da capacidade. Normalmente é usado o volume da 30ª ou da 50ª hora com maior volume.

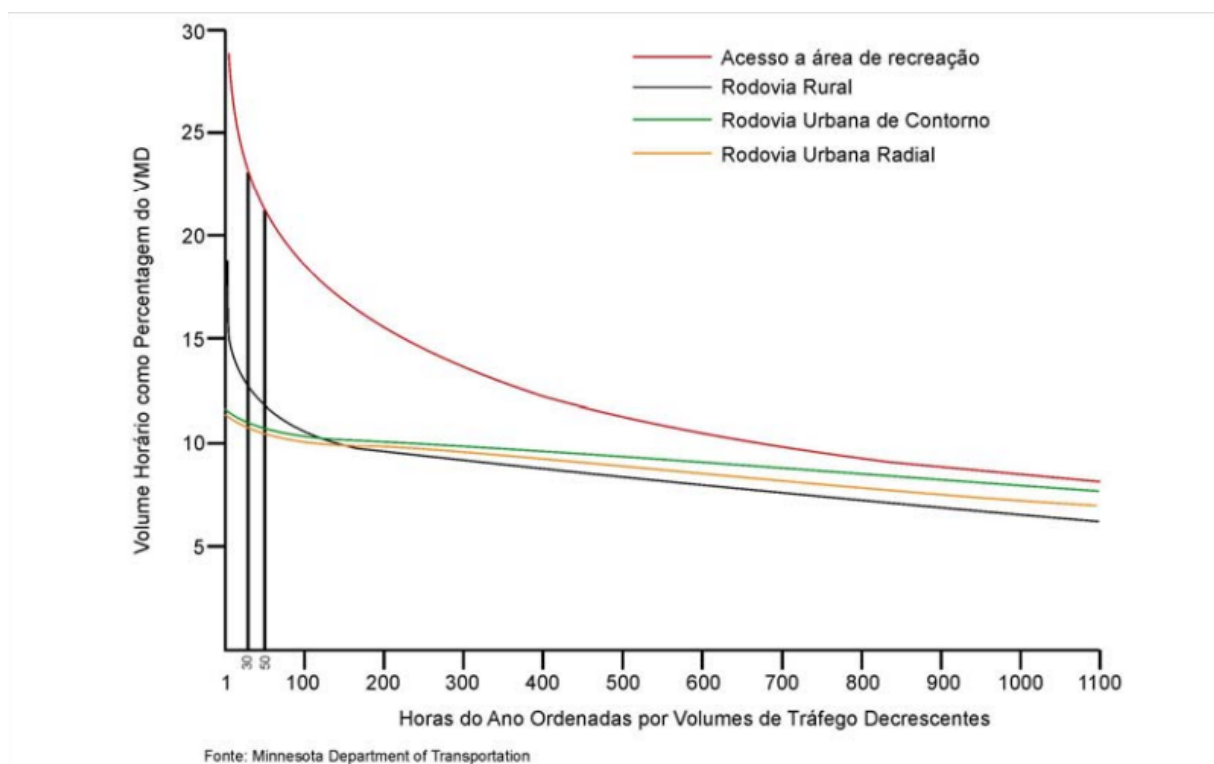
A curva de VHP é exemplificada pela Figura 1.

Na Figura 1 são observadas as curvas de volume horário, expressas em percentagem do VMD_a . E tem curvas para cada um dos 4 grupos de rodovias representadas quanto a classificação funcional das rodovias chamadas 'Acesso a áreas de recreação', 'Rodovia Rural', 'Rodovia Urbana de Contorno' e 'Rodovia Urbana Radial'. Isso acontece porque cada grupo de rodovias listado tem utilidades diferentes e a observação no gráfico deve ser feita de forma individual e a hora usada para determinar o VHP pode ser diferente.

As rodovias podem ser classificadas em urbanas ou rurais. Rodovias urbanas estão inseridas em áreas urbanizadas e costumam ser adaptações de caminhos antigos (sem planejamento de engenharia) atendendo principalmente aos deslocamentos diários casa-trabalho. Já rodovias rurais possuem alto padrão geométrico e são projetadas para conectar núcleos urbanos distantes, atravessando áreas não urbanizadas com controle de acesso. Mesmo quando cruzam áreas urbanas devido ao crescimento das cidades, mantêm sua função arterial.

Esta revisão se atém às rodovias rurais por ser onde estão localizada os postos de monitoramento usados na Seção 4 de aplicação.

Figura 1: Relação entre a hora e o volume horário de tráfego em rodovias norte-americanas (DNIT/IPR, 2006)



A caracterização da demanda de tráfego em rodovias rurais exige a análise integrada de fatores físicos, geográficos, funcionais e socioeconômicos. Por ligarem centros urbanos e atravessarem áreas não urbanizadas, essas vias apresentam padrões de tráfego heterogêneos, com significativa presença de veículos de carga e variações sazonais — como aumentos de volume durante períodos de colheita ou em épocas turísticas (PAPACOSTAS; PREVEDOUROS, 2009). Essa demanda é influenciada por elementos como o crescimento populacional, as atividades econômicas locais, a infraestrutura disponível e as características naturais do território (ORTÚZAR; WILLUMSEN, 2011). Compreender essa dinâmica é essencial para o planejamento eficiente da malha viária, priorização de investimentos e formulação de políticas públicas baseadas em modelos preditivos e medidas de segurança (DNIT/IPR, 2006).

A seguir, destacam-se os principais fatores geográficos e ambientais que impactam a geração e atração de tráfego nas rodovias rurais.

- **Fatores geográficos e ambientais:** elementos como relevo, cobertura vegetal, clima, presença de corpos hídricos e qualidade do solo influenciam diretamente o traçado das rodovias, os custos de manutenção e, conseqüentemente, o padrão de tráfego. Em regiões com relevo acidentado, por exemplo, o tráfego de veículos pesados pode ser reduzido ou condicionado a certos horários (FHWA, 2022).

- **Sazonalidade da demanda:** é comum que rodovias rurais apresentem variações sazonais significativas de volume, segundo FHWA (2022), associadas a:
 - atividades econômicas (colheitas, períodos de safra e escoamento da produção agrícola);
 - feriados e temporadas turísticas, especialmente em rotas de acesso a áreas de lazer, como mostrado na curva da Figura 1;
 - condições climáticas adversas (chuvas intensas ou estiagens) que podem influenciar o uso de certos trechos, principalmente aqueles não pavimentados.
- **Geração e atração de viagens:** rodovias rurais geralmente se conectam a zonas de geração de tráfego (áreas residenciais ou centros produtivos) e zonas de atração (centros urbanos, polos industriais, áreas de turismo ou centros de distribuição). A identificação desses polos permite estimar o potencial de tráfego futuro e fundamentar decisões de engenharia, conforme DNIT/IPR (2006).

Além disso, é importante reconhecer que a caracterização funcional da rodovia — por exemplo, se atua como via arterial de longa distância ou como acesso a propriedades rurais — influencia diretamente os padrões de tráfego e a criticidade de sua operação. Uma via de escoamento agrícola, por exemplo, pode ter tráfego concentrado em curtos períodos do ano, mas com volumes que justificam investimentos significativos (PAPACOSTAS; PREVEDOUROS, 2009). Em DNIT/IPR (2006) dados sócio-econômicos são tratados como informações básicas para caracterização do tráfego, com objetivo de "estabelecer relações entre os números de viagens realizadas e variáveis sócio-econômicas que possam explicá-las, de modo a possibilitar a determinação dos desejos de deslocamentos no futuro".

2.1.1 Contagem de tráfego

No Brasil, a principal contagem de tráfego é coordenada pelo DNIT, por meio do Plano Nacional de Contagem de Tráfego (PNCT), que coleta dados sistemáticos em pontos estratégicos da malha rodoviária federal, cujas localizações são definidas a fim de maximizar a abrangência do volume de tráfego monitorado, de acordo com PNCT (2023)¹. Esses dados subsidiam a tomada de decisão em relação à manutenção, à ampliação de capacidade e à segurança viária. Segundo a CNT (2016), os números obtidos são considerados essenciais para o planejamento de ações de conservação e desenvolvimento da infraestrutura rodoviária nacional.

Diversos métodos são utilizados para realizar a contagem de tráfego, cada qual com suas aplicações específicas e limitações. Alguns dos principais são apresentados a seguir.

¹site disponível em: <<https://servicos.dnit.gov.br/dadospnct/metodologia>>.

- **Contagem manual por observador humano:** realizada in loco por técnicos treinados que registram manualmente o número e o tipo de veículos que passam por uma seção de rodovia. Embora simples e de baixo custo, esse método é limitado em duração e sujeito a erros humanos (DNIT/IPR, 2006).
- **Sensores de pulso ou tubo pneumático:** instalam-se tubos flexíveis sobre a pista que, ao serem comprimidos pelas rodas dos veículos, geram pulsos de ar que são convertidos em dados de contagem. São eficazes para contagens temporárias e classificam os veículos com base no espaçamento entre eixos (PAPACOSTAS; PREVEDOUROS, 2009).
- **Laços indutivos e sensores piezoelétricos:** instalados permanentemente no pavimento, os laços indutivos detectam alterações na indutância (mantida constante por uma tensão elétrica aplicada) ocasionada pela passagem dos veículos metálicos, segundo Treiber e Kesting (2013). Enquanto os sensores piezoelétricos detectam a passagem de um eixo de veículo por meio de uma variação na voltagem do sensor, que é diretamente proporcional à pressão aplicada pela roda do veículo sobre o sensor, permitindo estimar o peso por eixo e o peso bruto total, além de captar a classificação veicular e a velocidade. São métodos altamente precisos e confiáveis, recomendados para medições contínuas e de longo prazo na contagem de tráfego e balanças (DNIT, 2018b; FHWA, 2022).
- **Câmeras e sistemas de processamento de imagem:** utilizam tecnologias de visão computacional para identificar, classificar e contar veículos. Têm a vantagem de coletar dados visuais ricos e de forma automatizada, embora exijam condições adequadas de iluminação e posição da câmera (DNIT/IPR, 2006).

O PNCT, segundo DNIT (2018b)², realiza contagem de tráfego com postos de contagem contínua e postos de cobertura. E os métodos de contagem usados são explicados a seguir.

- a) **Contagem contínua:** utiliza equipamentos fixos instalados em pontos estratégicos da malha rodoviária, operando 24 horas por dia, durante os 365 dias do ano. Esses dispositivos consistem em um receptor que emprega **laços indutivos e sensores piezoelétricos**, a Figura 2 ilustra a instalação de laços indutivos e a Figura 3 apresenta o esquema completo explicado. O esquema de equipamentos realiza a coleta de dados sobre o volume de tráfego, classificação dos veículos, velocidade operacional e peso (peso bruto total e por eixo). Os dados coletados são essenciais para analisar variações temporais e calcular indicadores como o Volume Médio Diário anual (VMD_a) e o Número N, que quantifica o número de repetições de eixo padrão e é fundamental para o dimensionamento de pavimentos.

²apresentação disponível em <<https://www.gov.br/dnit/pt-br/assuntos/campanhas-e-eventos/semana-do-planejamento/2a-semana-do-planejamento/2-divulgacao-via-web-dados-pnct-leonel-teixeira.pdf>>.

Figura 2: Equipamento contador em postos de coleta contínua do PNCT (DNIT, 2018b)

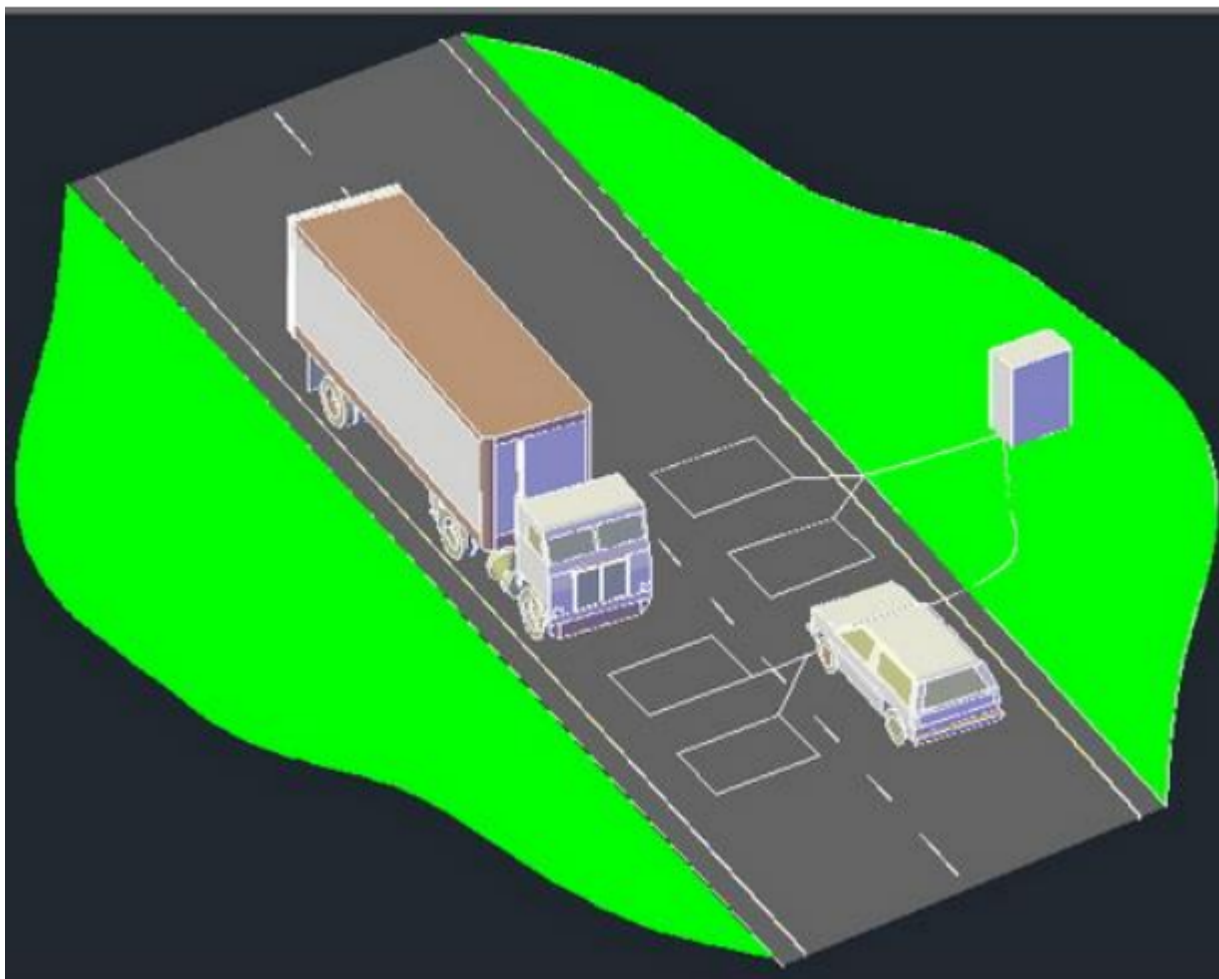
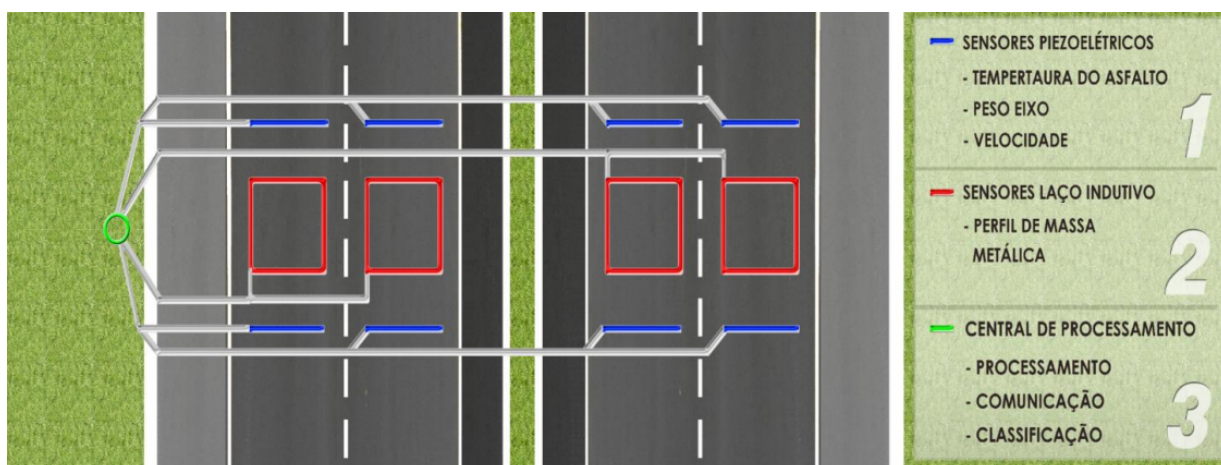


Figura 3: Explicação do equipamento contador em postos de coleta contínua do PNCT (DNIT, 2018b)



- b) **Contagem de cobertura:** realizada por meio de equipamentos portáteis e não intrusivos durante um período de 7 dias corridos em trechos definidos. Esses dispositivos utilizam a **técnica pneumática**, registrando volume, velocidade e classificação veicular. O objetivo dessa modalidade é ampliar a cobertura geográfica das contagens contínuas, oferecendo uma visão mais ampla das variações espaciais no tráfego.

Essas informações são integradas a modelos matemáticos de rede e a matrizes origem-destino, permitindo estimativas de tráfego para todo o Sistema Nacional de Vias (SNV) e auxiliando a definição de prioridades para investimentos e manutenção. As notas técnicas detalhadas da metodologia encontram-se disponíveis no portal do PNCT (DNIT, 2018a)³.

2.1.2 Expansão volumétrica

A expansão volumétrica é uma técnica utilizada para estimar o volume de tráfego em trechos da malha viária onde não há medição direta, com base em dados observados de outros postos de contagem. Trata-se de uma etapa essencial no planejamento de transportes, permitindo a ampliação da cobertura espacial das contagens de tráfego.

Esse processo consiste em utilizar os dados registrados em postos com coleta contínua ou de cobertura para inferir volumes de tráfego em locais não monitorados, com o intuito de alimentar modelos de transporte, subsidiar projetos de engenharia rodoviária e orientar decisões de gestão. A precisão das estimativas depende da escolha adequada dos postos de referência e da compreensão das variações temporais do tráfego. A expansão nunca será perfeita e sempre é preferível a contagem, mas DNIT/IPR (2006) descreve os resultados práticos de expansões volumétricas como aceitáveis e que justificam-se pela economia financeira gerada.

Para aplicar a expansão volumétrica entre dois postos, é fundamental estabelecer um grau de similaridade entre eles. A forma mais robusta de se estabelecer essa relação é por meio da análise comparativa das curvas de variação de tráfego ao longo do tempo, conforme indicado por Pignataro (1973). Quando curvas de variação horária, diária e sazonal se mostram similares entre dois postos, é possível aplicar fatores de expansão com maior confiabilidade.

Para compreender adequadamente a aplicação da expansão volumétrica, é necessário conhecer os diferentes tipos de variação de volume observados em sistemas rodoviários.

- **Variação horária (diurna):** O volume horário (Seção 2.1) varia ao longo do dia, com picos durante os horários de maior demanda (ex.: início da manhã e fim da tarde). Na Figura 4 a variação do volume horário é apresentada para os grupos de funcionalidade das rodovias americanas. E é feita a divisão por: quarta-feira (para representar os dias de semana), sábado

³apresentação disponível em <<https://www.gov.br/dnit/pt-br/assuntos/campanhas-e-eventos/semana-do-planejamento/2a-semana-do-planejamento/4-estimativa-vm-da-glydston-saul.pdf>>.

e domingo. A análise das curvas leva a crer que os horários de pico são mais bem definidos em vias locais. O entendimento das horas de pico é fundamental pois é nesse horário que ocorrem os eventos mais importantes, como acidentes e congestionamento.

O manual ainda traz a variação do volume dentro da hora de pico que estabelece 'Fator Horário de Pico (FHP)' que mede o grau uniformidade do fluxo através da equação (1) (DNIT/IPR, 2006).

$$FHP = \frac{V_{hp}}{4V_{15\max}} \quad (1)$$

onde:

- FHP = fator horário de pico
- V_{hp} = volume da hora de pico
- $V_{15\max}$ = volume do período de quinze minutos com maior fluxo de tráfego dentro da hora de pico

Valores altos indicam uniformidade entre os subintervalos de 15 minutos dentro da hora de pico e geralmente indicam grande volume de tráfego, podendo gerar congestionamento.

- **Variação semanal:** Refere-se à diferença de volume diário entre os dias da semana. Segundo DNIT/IPR (2006) as vias urbanas tem volumes maiores nos dias de semana, de segunda a sexta-feira e as rodovias rurais e rodovias de acesso a áreas de recreio têm volumes de pico de sexta-feira a domingo. Assume-se que as semanas de um mesmo mês são similares entre si.
- **Variação mensal (sazonal):** Representa as diferenças no volume médio de tráfego ao longo dos meses do ano. A variação sazonal é função do tipo de via e as atividades que ela serve (DNIT/IPR, 2006). Pode refletir fatores como estações do ano, calendário escolar, feriados e ciclos econômicos. A Figura 5 evidencia a diferença do impacto da sazonalidade, de forma geral, entre rodovias rurais e urbanas. As urbanas, que servem ao deslocamento ao trabalho, tendem a ter fluxo mais permanente, já as rurais são impactadas pelas férias escolares e período de turismo por serem também vias de acesso à áreas de recreação e turismo.
- **Variação anual:** Indica as mudanças no volume de tráfego de um ano para outro, influenciada por crescimento populacional, mudanças econômicas ou ampliação da infraestrutura. É necessário cuidado ao usar dados antigos por eles possivelmente levarem a uma avaliação errônea da importância real das rodovias (DNIT/IPR, 2006).
- **Outras variações:** Incluem variações por sentido de tráfego (ida/volta), por faixa de rolamento, e variações especiais provocadas por eventos atípicos como shows, acidentes, jogos de futebol ou feriados prolongados (DNIT/IPR, 2006).

Figura 4: Variações horárias típicas em vias americanas (DNIT/IPR, 2006)

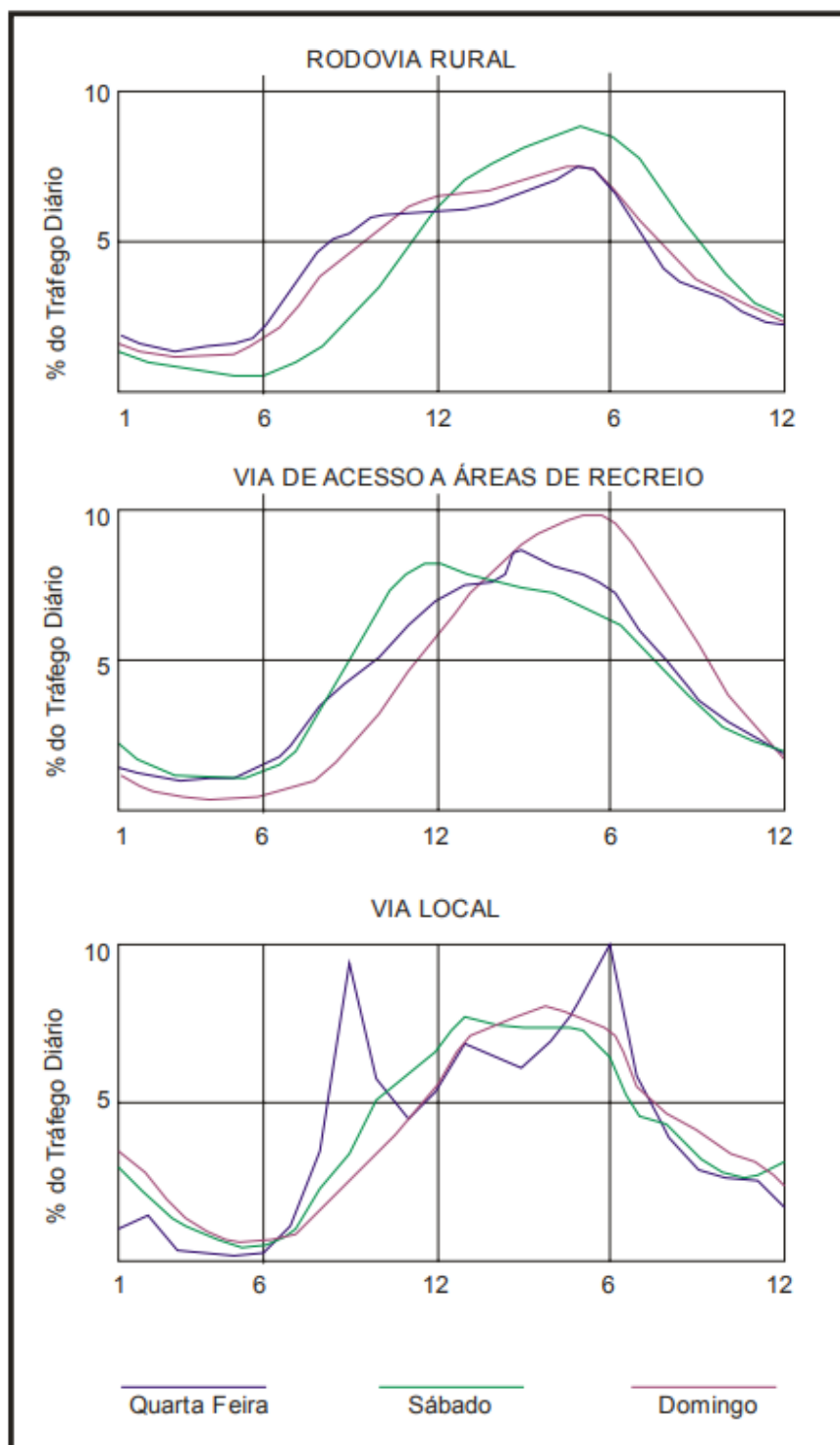
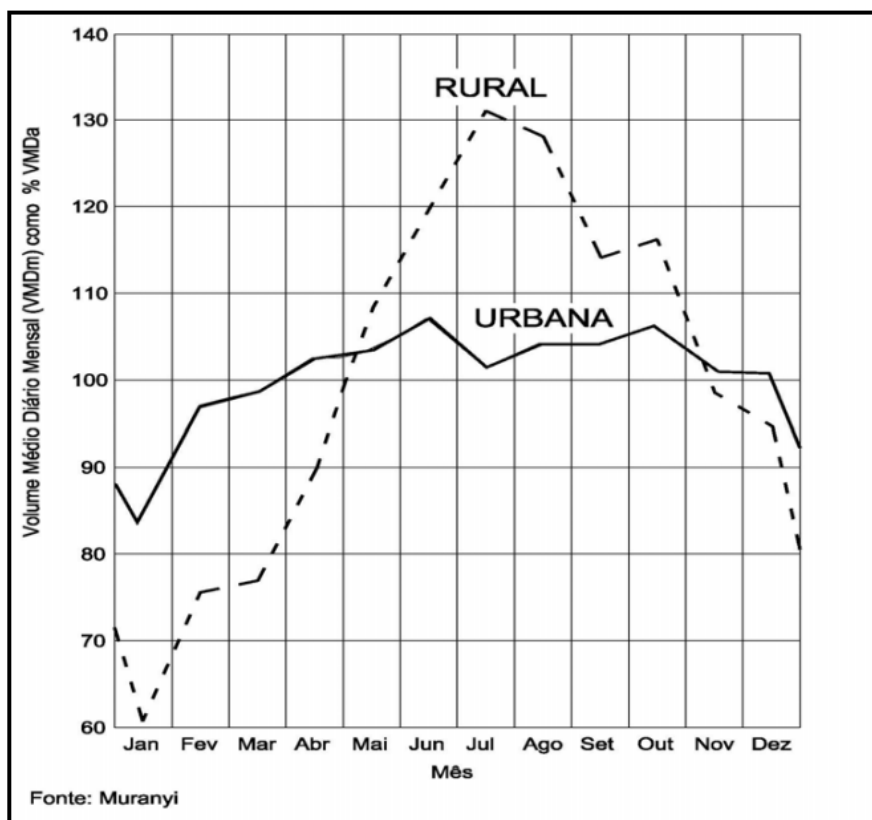


Figura 5: Variação mensal em rodovias rurais e urbanas (DNIT/IPR, 2006)



Para operacionalizar a expansão volumétrica, são utilizados os **fatores de expansão**, definidos como coeficientes aplicados a volumes parciais para estimar o volume total de um período maior. Cada fator é sempre calculado para cada tipo de veículo, sendo assim os valores usados nas fórmulas são do total observado do determinado tipo e os valores parciais também são relativos ao tipo de veículo sobre o qual se está expandindo o volume. A sequência lógica da expansão se dá conforme a Figura 6.

Segundo (DNIT/IPR, 2006), os fatores de expansão são os listados a seguir.

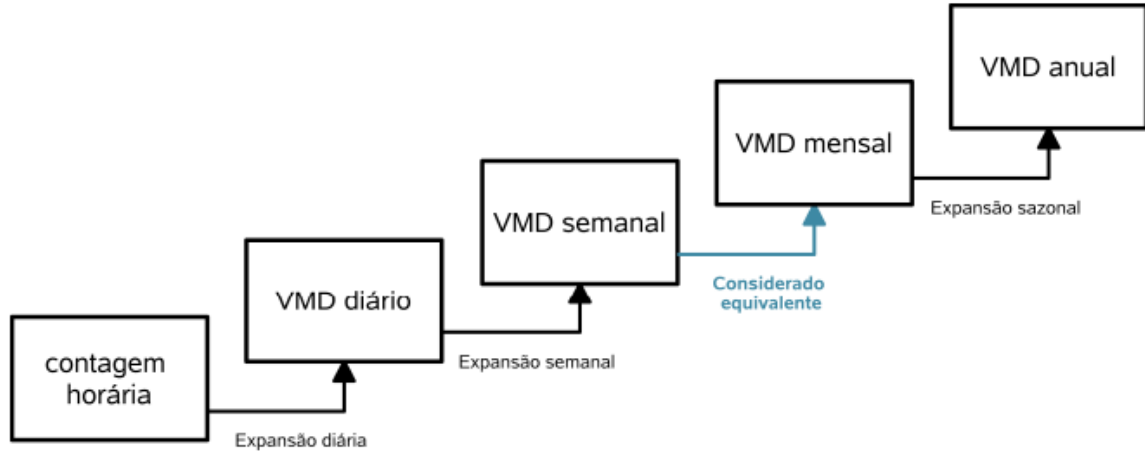
- **Fator de expansão diário:** permite extrapolar os dados de períodos reduzidos de contagem (ex.: 4h ou 12h) para o volume total do dia. Para cada hora do dia é atribuído um fator de expansão diário (para que se infira o volume total do dia com base no volume observado pelo posto que fez contagem em todas as horas do dia) esse fator é obtido pela equação (2).

$$f_d = \frac{VMD_d}{VH_h} \quad (2)$$

onde:

– f_d = fator de expansão diária;

Figura 6: Aplicação de fatores de expansão para transformar VMD horário, semanal e mensal (sazonal) para VMD_a (DNIT, 2008)



- $VMDd$ = volume de carros de passeio durante as 24 horas do dia;
- VH_h = volume de carros de passeio durante as h horas de pesquisa.

- **Fator de expansão semanal:** aplica-se para estimar o volume de tráfego de uma semana completa a partir de dados de um ou poucos dias. É feita a atribuição de fatores específicos para cada dia da semana, através da equação (3).

$$f_s = \frac{VMS}{VMDd} \quad (3)$$

onde:

- f_s = fator de expansão semanal de um dia específico da semana;
- VMS = volume de carros de passeio durante os 7 dias da semana;
- $VMDd$ = volume de carros de passeio durante o dia determinado.

Obtendo o $VMD_{semanal}$ e considerando que não há variação de volume médio diário semanal dentro de um mês, é possível aproximar $VMD_{semanal} \approx VMD_{mensal}$, que é usado para expansão mensal/sazonal.

- **Fator de expansão mensal/sazonal:** os fatores de expansão anual, se possível, devem ser determinados por zona de tráfego, uma vez que dependem da atividade econômica, que é variável por região. Os fatores de ajustamento sazonal são obtidos para cada tipo de veículo pela equação 4.

$$f_a = \frac{VMD_a}{VMD_m} \quad (4)$$

onde:

- f_a = fator de expansão sazonal;
- VMD_a = volume médio diário anual;
- VMD_m = volume médio diário mensal.

Um posto de contagem contínua, que tenha realizado contagem de dados 24h por dia durante 365 dias consecutivos de um ano, e tem, portanto, o VMD_a obtido por observação, tem capacidade de ser usado para calcular qualquer um dos fatores de expansão para outro posto, contando, claro, que seja determinada a similaridade entre os postos.

2.2 Estabelecimento de relações entre postos

Para a engenharia de tráfego, estabelecer relações entre diferentes locais de medição é uma prática essencial. Essas relações permitem compreender padrões de variação de volume ao longo do tempo e do espaço, possibilitando o uso de dados de uma estação em outra com características semelhantes. Além da estimativa de volumes em locais não monitorados, essa lógica também pode servir como base para a expansão volumétrica de contagens amostrais, redução de custos operacionais e inferência de padrões em regiões com infraestrutura limitada.

No contexto da estimativa do volume médio diário anual (VMD_a), as relações entre postos de contagem são utilizadas para converter dados obtidos em contagens de curto prazo em estimativas anuais representativas, conforme visto na revisão de expansão volumétrica, Seção 2.1.2 segundo DNIT/IPR (2006), em sua explicação, faz uso exclusivo de curvas de variação volumétrica para definir similaridade entre os postos. Nesta seção, é apresentado um novo contexto através de uma revisão de fontes americanas, primariamente, e é importante perceber como elas incluem variáveis geográficas e variáveis não numéricas.

Neste descrito contexto, a metodologia tradicional contempla cinco passos principais para estimar VMD_a , conforme a revisão realizada por Schneider e Tsapakis (2009) do Departamento de Transportes de Ohio-EUA, em alinhamento com as diretrizes do Traffic Monitoring Guide (TMG) da Federal Highway Administration (FHWA):

- **Passo 1:** Coletar e limpar os dados de volume e classificação de tráfego a partir de postos de contagem contínua;
- **Passo 2:** Estimar os fatores de ajuste (AF) ou fatores de ajuste sazonais (SAF), que podem incluir hora do dia, dia da semana, mês/estação do ano;

- **Passo 3:** Agrupar os dados de tráfego de acordo com seus padrões de tráfego semelhantes e com base nos fatores de ajuste (AFs) ou fatores de ajuste sazonal (SAFs).
- **Passo 4:** Associar postos ou segmentos viários onde são feitas contagens de curto prazo com os grupos formados a partir das contagens contínuas;
- **Passo 5:** Aplicar os fatores sazonais (SAFs) calculados no Passo 2 a cada contagem de curto prazo, a fim de obter a estimativa de VMD_a para aquele local.

Com o intuito de resolver o Passo 3, existem três abordagens principais para agrupar SAFs: (i) atribuição geográfica ou funcional de rodovias a grupos, (ii) análise de agrupamento estatístico (*cluster analysis*) e (iii) aplicação direta de fatores.

(i) Pesquisas diversas já chegaram a alguns conjuntos de grupos baseado em dados geográficos/funcionais. Bellamy (1978), por exemplo, classificou postos de monitoramento em quatro grupos: recreacional, não recreacional de baixo fluxo, rural de longa distância e urbano de trajeto diário (ou urbano pendular). Segundo Drusch (1966) um critério de diferença de 0,20 entre os valores máximo e mínimo dos Fatores de Ajuste Sazonal (SAFs) dentro de cada mês, para as rodovias (ou localidade dos postos) de um mesmo grupo, é considerado aceitável. A atribuição geográfica ou funcional, embora amplamente utilizada, depende fortemente de julgamento técnico, assumindo que vias com a mesma classificação funcional compartilham padrões sazonais semelhantes — o que nem sempre é verdade empiricamente.

(ii) A análise de agrupamento estatístico surge como uma alternativa orientada por dados. Ela utiliza algoritmos para identificar padrões de variação de volume ao longo do tempo, agrupando postos com comportamento semelhante. Um ponto positivo dessa abordagem é que reduz o viés do julgamento técnico e pode revelar padrões que não são evidentes. Contudo, a principal desvantagem desse método é que a lógica que gera os resultados dos grupos pode ser difícil de interpretar, especialmente quando eles não se alinham a categorias funcionais tradicionais.

Estudos em Shah e Hirschmann (1981) mostram que as estimativas de VMD_a podem ser significativamente melhoradas quando variáveis explicativas adicionais são consideradas no processo de agrupamento. Isso é especialmente relevante para este trabalho, que busca empregar múltiplas variáveis geográficas/sócio-econômicas na clusterização de postos.

Além da clusterização, Schneider e Tsapakis (2009) também apresentam métodos avançados como análise discriminante, árvores de decisão com lógica *fuzzy*, vizinho mais próximo e outros modelos baseados em inteligência artificial. Esses métodos têm como objetivo melhorar a atribuição de contagens de curto prazo a grupos de SAFs adequados, mas sua eficácia depende diretamente da qualidade e profundidade das variáveis utilizadas.

A versão mais recente do TMG (FHWA, 2022) faz uso de fatores de grupos para ajustar contagens de volume de curta duração para VMD_a (expansão volumétrica). A contagem de volume

de curto prazo geralmente requer diversos ajustes para converter uma contagem bruta de volume diário de tráfego em uma estimativa de VMD_a . O conjunto específico de ajustes necessários depende do equipamento utilizado para realizar a contagem e da própria duração da contagem. Quase todas as contagens de curto prazo exigem ajustes para reduzir os efeitos de viés temporal, caso essas contagens sejam utilizadas para estimar o VMD_a . Em geral, uma contagem de eixos de 48 horas é convertida para VMD_a com a equação (5):

$$VMD_{ahi} = VOL_{hi} \times M_h \times D_h \times T_h \times A_i \times G_h \quad (5)$$

Onde:

- VMD_{ahi} = volume médio diário anual na localização i do grupo de fator h ;
- VOL_{hi} = volume de eixos de 48 horas na localização i do grupo de fator h ;
- M_h = fator mensal (sazonal) aplicável para o grupo de fator h ;
- D_h = fator aplicável para o dia da semana (DOW - *day of week*) do grupo de fator h (se necessário);
- T_h = fator aplicável para a hora do dia (TOD - *time of day*) do grupo de fator h (se necessário, para contagens de dias parciais);
- A_i = fator aplicável de correção de eixos para a localização i (caso não seja uma contagem de volume ou classe de tráfego, por exemplo, contagens coletadas com tubo pneumático simples);
- G_h = fator aplicável de variação anual (por exemplo, taxa de crescimento ou declínio) para o grupo de fator h (se necessário).

Essa fórmula deve ser modificada conforme o tipo de equipamento de coleta, duração da contagem e ano de referência. Por exemplo, se a contagem for realizada com detector por laço indutivo em vez de tubo pneumático, remove-se o fator A_i . Se a contagem abranger sete dias consecutivos, os volumes podem ser agregados, e o fator D_h é eliminado. Pequenas adaptações, como incluir uma 49ª hora em uma contagem de 48h, podem melhorar a acurácia do resultado final.

A criação de grupos de fatores de ajuste (hora do dia, dia da semana e mês do ano) por tipo de veículo ainda é um campo em desenvolvimento em muitos estados americanos (FHWA, 2022). Para isso, o guia propõe as seguintes quatro abordagens principais.

- **Agrupamento baseado em conhecimento (*knowledge-based grouping*)**: utiliza dados disponíveis e o conhecimento prático de especialistas para identificar padrões de tráfego. Considera fatores como atividades econômicas locais, presença de tráfego de passagem, uso do solo e características funcionais da via. Rodovias com indústrias, áreas agrícolas ou rotas

turísticas, por exemplo, apresentam padrões distintos. O grupo pode ser ajustado conforme novas informações indicarem que uma via se comporta de maneira diferente.

- **Análise de agrupamento estatístico (*cluster analysis*):** aplica métodos estatísticos para identificar padrões naturais de variação no tráfego, com base em dados de volume por classe veicular. Permite agrupar postos de contagem contínua de forma objetiva e determinar quantos grupos são apropriados.
- **Área de influência:** considera o contexto geográfico e o tipo de atividades predominantes na região. As vias são agrupadas conforme sua localização e influência regional (ex: áreas industriais, agrícolas, urbanas ou turísticas).
- **Abordagem híbrida:** combina o agrupamento baseado em conhecimento com a análise estatística. Une a objetividade dos dados com a interpretação prática de especialistas, resultando em grupos bem definidos e aplicáveis.

Em resumo, a literatura revisada por Schneider e Tsapakis (2009) e o TMG (FHWA, 2022) reforçam que a incorporação de variáveis explicativas, pode elevar significativamente a precisão das estimativas de VMD_a . Essa conclusão sustenta a proposta deste trabalho, que utiliza técnicas de clusterização para agrupar postos de monitoramento de tráfego baseado em múltiplas características geográficas.

2.3 Aprendizado de máquinas

Segundo Russell e Norvig (2020, p. 2), “aos olhos do público, às vezes há confusão entre os termos ‘inteligência artificial (IA)’ e ‘aprendizado de máquina’”. O aprendizado de máquina é um subcampo da IA que estuda a capacidade de melhorar o desempenho com base na experiência.” Essa capacidade de adaptação torna o aprendizado de máquina essencial em contextos onde há grande volume de dados e variabilidade de padrões, como em reconhecimento de imagens, diagnósticos médicos, previsão de demanda, entre outros.

As origens do aprendizado de máquina remontam à década de 1940, quando Warren McCulloch e Walter Pitts propuseram um modelo de neurônio artificial no artigo “*A Logical Calculus of the Ideas Immanent in Nervous Activity*” (MCCULLOCH; PITTS, 1943). Eles demonstraram que redes de neurônios artificiais poderiam realizar operações lógicas, antecipando os conceitos que viriam a fundamentar as redes neurais modernas. Poucos anos depois, Alan Turing contribuiu para o debate com o artigo “*Computing Machinery and Intelligence*” (TURING, 1950), no qual propôs o famoso Teste de Turing. Essa obra seminal introduziu questionamentos sobre a capacidade das máquinas de simular o raciocínio humano, abrindo caminho para o desenvolvimento de sistemas capazes de aprender com dados.

O aprendizado de máquina surgiu como alternativa aos sistemas baseados em regras manuais, nos quais o programador define todas as condições de decisão. Conforme (MÜLLER; GUIDO, 2016), este modelo rígido apresenta limitações:

- a lógica de decisão é específica para uma tarefa. Pequenas mudanças no domínio exigem reprogramação completa.
- criar regras exige conhecimento detalhado do processo decisório, o que nem sempre está disponível.

Um exemplo que ilustra essas limitações é a detecção de rostos em imagens digitais. A percepção humana de um rosto não se traduz facilmente em regras para computadores, mas algoritmos de aprendizado de máquina podem identificar tais padrões com base em grandes conjuntos de imagens rotuladas (BISHOP, 2006).

Para entender as principais abordagens de aprendizado de máquina, é necessário distinguir entre suas categorias e conceitos fundamentais.

Tipos de aprendizado.

- **Aprendizado supervisionado:** Segundo Jolly (2018), aprendizado supervisionado é uma forma de aprendizado de máquina em que nossos dados vêm acompanhados de um conjunto de rótulos ou uma variável-alvo numérica. Esses rótulos/categorias geralmente pertencem a um único atributo/característica, comumente conhecido como variável-alvo.
- **Aprendizado não supervisionado:** É uma forma de aprendizado de máquina em que o algoritmo tenta detectar/encontrar padrões em dados que não possuem uma variável de resultado ou variável-alvo. Em outras palavras, não temos dados com rótulos preexistentes. Assim, o algoritmo normalmente utiliza uma métrica como distância para agrupar os dados com base na proximidade entre si.

O autor exemplifica: cada linha dos seus dados pode pertencer à categoria "Saudável" ou "Não Saudável". Dado um conjunto de variáveis como peso, níveis de açúcar no sangue e idade, podemos usar um algoritmo de aprendizado supervisionado para prever se a pessoa é saudável ou não. O algoritmo de aprendizado não supervisionado, para o mesmo problema, utilizará uma métrica, como a distância, para identificar o quão próximos estão os pontos entre si e o quão distantes estão dois grupos formados. Em seguida, o algoritmo procederá à clusterização desses grupos em dois conjuntos distintos (JOLLY, 2018). Caberia a uma pessoa a tarefa de julgar se os grupos formados respondem a pergunta e a qualidade do agrupamento estaria intimamente ligada a quão indicativas para a situação são as variáveis usadas.

Conceitos técnicos importantes.

- **Generalização:** se um modelo é capaz de fazer previsões precisas em dados não vistos, ele é considerado capaz de generalizar do conjunto de treinamento para o conjunto de teste. O objetivo é construir um modelo que consiga generalizar da forma mais precisa possível (MÜLLER; GUIDO, 2016).
- **Overfitting:** o modelo foi treinado de forma muito específica aos exemplos de treinamento fornecidos e não se generaliza bem para casos ou exemplos de dados que ele ainda não encontrou (JOLLY, 2018).
- **Underfitting:** o modelo se ajusta aos dados de maneira muito genérica e não apresenta bom desempenho na previsão dos rótulos corretos das classes no conjunto de teste (JOLLY, 2018).

Segundo Müller e Guido (2016), quanto mais complexo for um modelo, melhor ele conseguirá fazer previsões sobre os dados de treinamento. No entanto, se este modelo se tornar excessivamente complexo, ele começará a focar demais em cada ponto individual do conjunto de treinamento e não será capaz de generalizar bem para novos dados. Existe um ponto de equilíbrio (*sweet spot*) entre esses extremos que proporcionará o melhor desempenho de generalização — é esse o modelo que é buscado.

Principais técnicas de aprendizado de máquina e exemplos de aplicação.

- **Regressão Linear:** a regressão linear é utilizada para prever um valor numérico contínuo a partir de um conjunto de variáveis de entrada (JOLLY, 2018). Exemplo: estimar o consumo de energia de uma residência.
- **Classificação (ex: árvores de decisão):** em termos simples é um conjunto de regras que nos ajuda a classificar observações em grupos distintos (JOLLY, 2018). Exemplo: identificar se uma transação bancária é fraudulenta ou legítima.
- **Redes neurais:** modelos compostos por camadas de neurônios artificiais, eficazes em tarefas complexas, de muitas dimensões (JOLLY, 2018). Exemplo: sistemas de desbloqueio facial em *smartphones*.
- **Agrupamento:** agrupa dados semelhantes sem rótulos prévios, útil na identificação de padrões emergentes (JOLLY, 2018). Exemplo: detectar grupos de regiões com comportamento de tráfego similar.

2.4 Agrupamento

Agrupamento, segundo a classificação de Bishop (2006), é uma técnica de aprendizado de máquina não supervisionado cujo objetivo é identificar grupos de exemplos similares dentro dos dados, sem que estes grupos estejam previamente definidos.

De acordo com Han, Kamber e Pei (2006), o processo de agrupar um conjunto de objetos físicos ou abstratos em classes de objetos semelhantes é denominado *clustering*. Um *cluster* é um conjunto de pontos de dados que são similares entre si dentro do *cluster* e dissimilares em relação aos pontos de outros *clusters*. Um *cluster* pode ser tratado coletivamente como uma unidade em diversas aplicações.

De forma semelhante aos algoritmos de classificação, os algoritmos de agrupamento atribuem um número ou rótulo a cada ponto de dados, indicando a qual *cluster* esse ponto pertence.

Ainda segundo Han, Kamber e Pei (2006), a escolha do algoritmo de agrupamento depende dos tipos de dados disponíveis e do objetivo específico da aplicação. Os autores organizam os principais métodos de agrupamento nas seguintes cinco categorias.

- **Particionamento**

Dada uma base de dados com n objetos ou tuplas de dados, um método de particionamento constrói k partições dos dados, onde cada partição representa um *cluster*, e $k \leq n$. Ou seja, ele classifica os dados em k grupos, que juntos satisfazem os seguintes requisitos: (1) cada grupo deve conter pelo menos um objeto, e (2) cada objeto deve pertencer a exatamente um grupo. O segundo critério pode ser relaxado em abordagens fuzzy⁴.

A partir do número de partições k a serem construídas, um método de particionamento cria uma partição inicial. Ele então utiliza uma técnica de realocação iterativa que tenta melhorar a partição movendo objetos de um grupo para outro. O critério geral de uma boa partição é que os objetos no mesmo *cluster* estejam "próximos" ou relacionados entre si, enquanto os objetos de *clusters* diferentes estejam "distantes" ou sejam muito diferentes. Existem vários outros critérios para julgar a qualidade das partições.

Alcançar a otimização global em métodos de agrupamento baseados em particionamento exigiria a enumeração exaustiva de todas as partições possíveis. Em vez disso, a maioria das aplicações adota um dos dois métodos heurísticos populares: (1) o algoritmo *k-means*, onde cada *cluster* é representado pelo valor médio dos objetos no *cluster*; e (2) o algoritmo *k-medoids*, onde cada *cluster* é representado por um dos objetos localizados próximo ao centro do *cluster*. Esses métodos heurísticos de agrupamento funcionam bem para encontrar *clusters* de formato esférico em bases de dados pequenas a médias. Para encontrar *clusters* com formas complexas e para agrupamento de conjuntos de dados muito grandes, os métodos baseados em particionamento precisam ser estendidos.

- **Hierárquico**

⁴Nas abordagens fuzzy cada ponto tem um "grau de pertinência" (ou grau de associação) com cada *cluster*.

Métodos hierárquicos de clusterização constroem uma estrutura em forma de árvore, onde os agrupamentos são definidos de forma sequencial e aninhada. Essa estrutura pode seguir uma lógica **aglomerativa**, também chamada de “de baixo para cima” (*bottom-up*), na qual os dados começam isolados e sucessivamente são fundidos, ou **divisiva**, conhecida como “de cima para baixo” (*top-down*), onde se parte de um único grupo e ele desmembra-se. A escolha do caminho influencia diretamente na formação dos *clusters*, mas em ambos os casos as decisões tomadas não podem ser revertidas posteriormente — o que dá ao método sua simplicidade e rapidez, mas também sua limitação.

Apesar dessa rigidez, esses métodos funcionam bem em situações onde o custo computacional precisa ser controlado e onde há interesse em entender a relação hierárquica entre os dados. Para tornar a técnica mais flexível e precisa, é comum combiná-la com outros métodos, como a realocação iterativa. Abordagens modernas como BIRCH e CURE adotam essa lógica híbrida e conseguem resultados mais eficazes mesmo em grandes volumes de dados.

- Baseado em densidade

A maioria dos métodos de partição agrupa objetos com base na distância entre eles. Esses métodos geralmente conseguem identificar apenas *clusters* com forma esférica e apresentam dificuldades para descobrir agrupamentos de formatos arbitrários. Métodos de clusterização foram desenvolvidos com base na noção de densidade. A ideia geral é continuar expandindo um *cluster* enquanto a densidade (número de objetos ou pontos de dados) na “vizinhança” exceder um certo limiar. Ou seja, para cada ponto de dados dentro de um determinado *cluster*, a vizinhança com um raio definido deve conter pelo menos um número mínimo de pontos. Esse tipo de método pode ser utilizado para filtrar ruídos (*outliers*) e identificar *clusters* com formas arbitrárias.

O DBSCAN é um método típico baseado em densidade, que expande *clusters* de acordo com um limiar de densidade. O OPTICS é outro método baseado em densidade, que calcula uma ordenação aumentada de clusterização para análise interativa e automática de *clusters*.

- Baseado em grade

Um método baseado em grade quantiza o espaço dos objetos em um número finito de células que formam uma estrutura em grade. Ele então realiza todas as operações de clusterização sobre essa estrutura (isto é, no espaço quantizado).

A principal vantagem dessa abordagem é o seu tempo de processamento rápido, que normalmente é independente do número de objetos de dados e depende somente do número de células em cada dimensão do espaço quantizado.

STING é um exemplo típico de método baseado em grade. CLIQUE e Wave-Cluster são dois algoritmos de clusterização que são tanto baseados em grade quanto baseados em densidade.

- Baseado em modelo

Um método baseado em modelo propõe um modelo para cada um dos *clusters* e encontra o melhor ajuste dos dados a esse modelo. Um algoritmo baseado em modelo pode localizar *clusters* construindo uma função de densidade que reflete a distribuição espacial dos pontos de dados.

Esse método também permite determinar automaticamente o número de *clusters* com base em estatísticas padrão, levando em consideração ruídos ou *outliers* e, assim, gerando métodos de clusterização mais robustos.

Alguns algoritmos de clusterização integram ideias de vários métodos de clusterização, de forma que, às vezes, é difícil classificar um algoritmo como pertencente unicamente a uma única categoria de método. Além disso, algumas aplicações podem apresentar critérios de clusterização que exigem a integração de várias técnicas (HAN; KAMBER; PEI, 2006).

2.4.1 *K-means*

O objetivo de *K-means* é particionar o conjunto de dados em um número K de *clusters* (BISHOP, 2006). O algoritmo requer a definição deste valor. Dois métodos para definir o número ideal de *clusters* são o índice de silhueta e o método do cotovelo.

Para entender estes métodos, é necessário definir os dois seguintes conceitos.

1. **Distância euclidiana**, segundo Han, Kamber e Pei (2006).

Após a padronização, ou mesmo sem padronização, em certas aplicações, calcula-se a dissimilaridade (ou similaridade) entre os objetos. Quando se utilizam variáveis em escala intervalar, isso geralmente é feito com base na distância entre cada par de objetos. Existem algumas abordagens para definir essa distância. A medida de distância mais popular é a **distância euclidiana**, definida como:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2} \quad (6)$$

onde $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ são dois pontos de dados p -dimensionais.

2. **Soma das Diferenças Quadráticas (*Sum of Squared Errors*) - SSE** conforme Almeida, Carvalho e Menino (2017).

Para verificar o erro de cada grupo, é usada a soma das diferenças quadráticas entre cada ponto e o centroide do seu grupo. Para um grupo 1:

$$SSE(\text{grupo}_1) = \sum_{x_i \in \text{grupo}_1} \text{dist}(x_i - c_1)^2 \quad (7)$$

Em que x_i representa cada ponto do grupo 1, c_1 representa o centroide deste mesmo grupo e ‘dist’ a medida de distância utilizada (distância euclidiana) que pode ser escrita como na equação (8).

$$SSE(\text{grupo}_1) = \sum_{x_i \in \text{grupo}_1} \|x_i - c_1\|^2 \quad (8)$$

Então, para calcular o erro total dos grupos, basta fazer o somatório de SSE de cada grupo. No caso de haver 3 grupos:

$$SSE_{\text{total}} = SSE(\text{grupo}_1) + SSE(\text{grupo}_2) + SSE(\text{grupo}_3) \quad (9)$$

- **Método do Cotovelo**

O método do cotovelo usa o SSE total (de todos os grupos) em cada agrupamento para determinar o número ideal de *clusters* K, como apresentado na Figura 7. A ideia do método é gerar agrupamentos com diferentes quantidades de grupos. À medida que o número de *clusters* aumenta, o erro total tende a diminuir, aproximando-se de zero quando há um ponto por grupo, ou seja, quando cada observação forma seu próprio *cluster* (ALMEIDA; CARVALHO; MENINO, 2017).

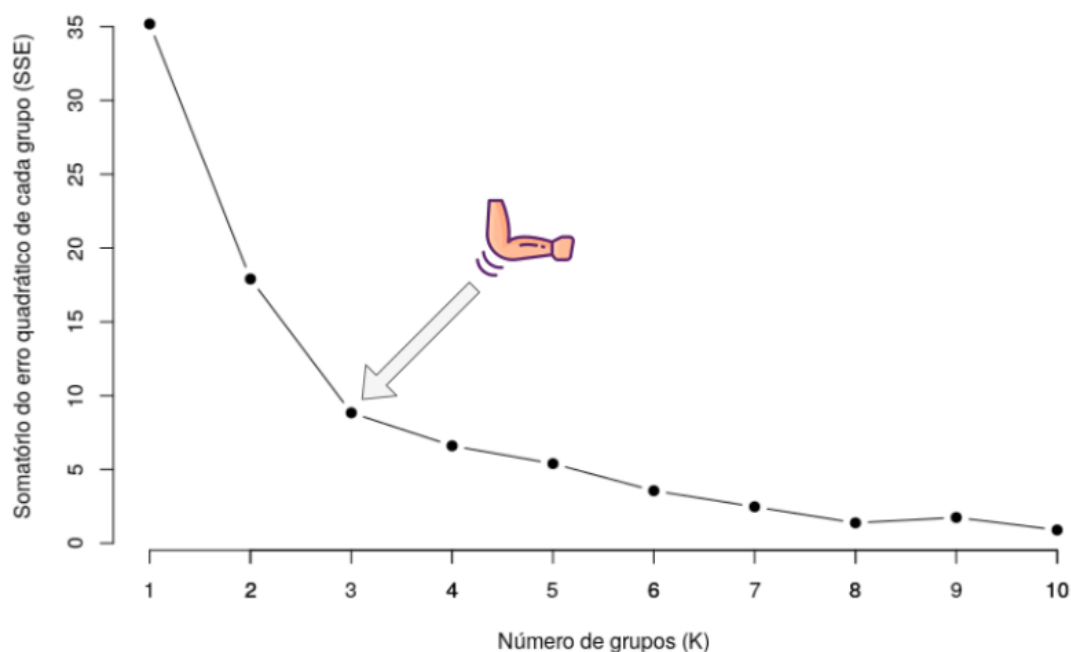
O número ideal de grupos se dá quando o ponto forma uma curva semelhante a um cotovelo. Este ponto é considerado ideal pois os pontos acima do cotovelo estão em uma região de *underfitting* e abaixo dele em *overfitting*, conceitos definidos na Seção 2.3 (ALMEIDA; CARVALHO; MENINO, 2017).

- **Índice de Silhueta**

Mede a semelhança de um ponto de dados dentro de seu *cluster* (coesão) em comparação com outros *clusters* (separação) (MICHILLES, 2025). O índice de Silhueta varia de -1 a 1. Valores próximos a 1 indicam que o objeto possui semelhança com objetos de seu grupo e dessemelhança com objetos de outros grupos (ALMEIDA; CARVALHO; MENINO, 2017).

O cálculo se dá da seguinte maneira:

Figura 7: Método do cotovelo em conjunto exemplo de dados (ALMEIDA; CARVALHO; MENINO, 2017)



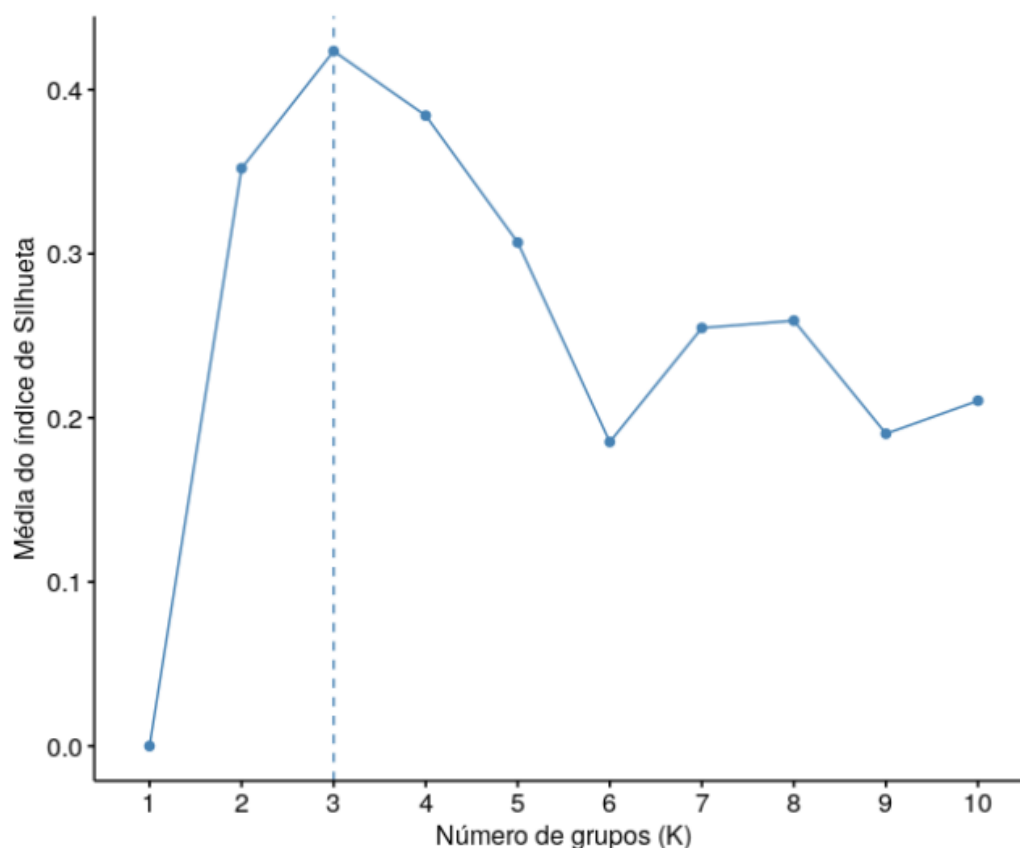
- para cada ponto individual i , os valores de coesão e separação que compõem as médias a e b , respectivamente, são calculados por:
 - * a coesão do ponto i dentro de um *cluster* é dada pela média da distância de i para todos os outros pontos dentro do mesmo *cluster*.
 - * a separação é calculada pela distância entre o ponto i e o ponto mais próximo de qualquer outro *cluster*.
- a fórmula do coeficiente de silhueta (MICHILLES, 2025) é dada por:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (10)$$

- * $S(i)$ é o coeficiente de silhueta do ponto i ;
- * $a(i)$ é a distância média entre i e todos os outros pontos de dados no *cluster* ao qual ele pertence;
- * $b(i)$ é a distância média de i a todos os *clusters* que ele não pertence.

Almeida, Carvalho e Menino (2017) apresentam a Figura 8, o gráfico é formado a partir de dados de exemplo.

Figura 8: Método da silhueta (ALMEIDA; CARVALHO; MENINO, 2017)



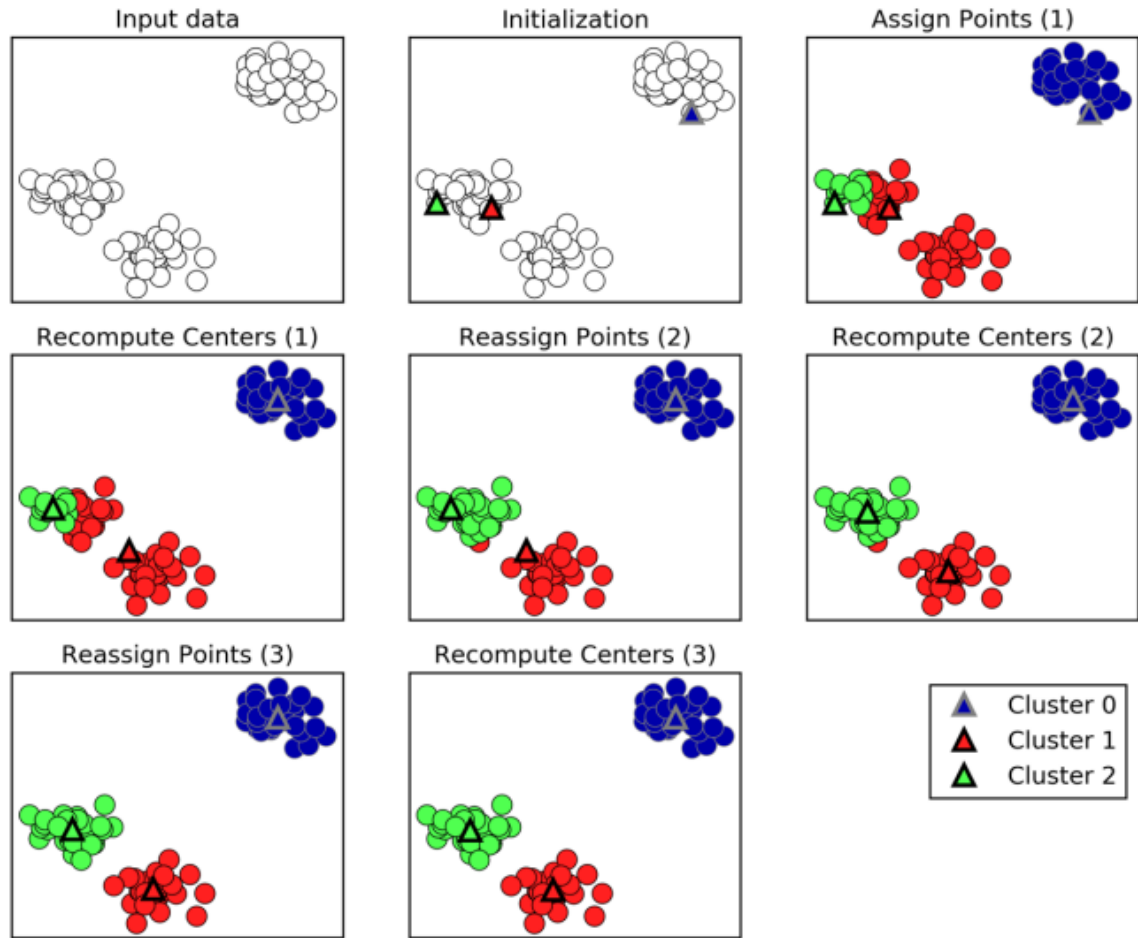
No exemplo da Figura 8 é recomendado usar 3 *clusters*, por ser a quantidade que apresenta a maior média do índice de silhueta (ALMEIDA; CARVALHO; MENINO, 2017).

Definido o número de *clusters*, o primeiro passo que o algoritmo realiza é atribuir um conjunto de centroides aleatórios (JOLLY, 2018).

A partir disso, inicia-se um *loop* em que o algoritmo alterna entre dois passos: atribuir cada ponto de dado ao centroide de *cluster* mais próximo e, em seguida, definir cada centroide como a média dos pontos de dado que lhe são atribuídos. O algoritmo é finalizado quando a atribuição dos pontos aos *clusters* não muda mais ou quando é atingido o número de repetições predefinido. A Figura 9 ilustra, com dados sintéticos, o processo feito pelo algoritmo. Mostrando os dados de entrada, posicionamento aleatório dos centroides (*initialization*), atribuição dos pontos ao centroide mais próximo, reposicionamento dos centroides na média de seus pontos e a repetição dessas etapas até atingir um dos requisitos de parada.

Os autores Han, Kamber e Pei (2006) resumem os pontos vantajosos e desvantajosos do algoritmo *K-means*: ele funciona bem quando os *clusters* são nuvens compactas que estão razoavelmente bem separadas umas das outras. O método é relativamente escalável e eficiente no processamento

Figura 9: Posicionamento dos centroides e atribuição dos pontos (MÜLLER; GUIDO, 2016)



de grandes conjuntos de dados, pois a complexidade computacional do algoritmo é $O(nkt)$, onde n é o número total de objetos, k é o número de *clusters*, e t é o número de iterações. Normalmente, $k \ll n$ e $t \ll n$. O método frequentemente termina em um ótimo local.

O método *K-means*, entretanto, só pode ser aplicado (diretamente) quando a média de um *cluster* está definida. Isso pode não ser o caso em algumas aplicações, como quando os dados possuem atributos categóricos. A necessidade de os usuários especificarem k , o número de *clusters*, antecipadamente pode ser vista como uma desvantagem. O método *K-means* não é adequado para descobrir *clusters* com formas não convexas ou de tamanhos muito diferentes. Além disso, é sensível a ruídos e pontos de dados discrepantes, já que um pequeno número desses dados pode influenciar substancialmente o valor da média.

3 Método

O fluxo de trabalho é linear e sequencial, ou seja, cada etapa depende da conclusão da anterior para sua execução e elas seguem a seguinte ordem.

1. **Extração de dados dos postos de monitoramento de tráfego** - dados de identificação e localização.
2. **Aquisição de dados geográficos** - dados descritivos dos postos ou da região que caracterizem o posto.
3. **Preparação dos dados para o agrupamento** - tratamento de valores nulos e codificação de variáveis categóricas.
4. **Agrupamento dos postos através do algoritmo *K-means*** - normalização dos dados; seleção das colunas a serem usadas para agrupar; definição do número de *clusters* e execução do agrupamento.
5. **Calibração do modelo** - verificação dos resultados; ajustes no algoritmo *K-means*; visualização dos resultados finais.
6. **Relacionamento dos postos de cobertura ao posto permanente mais similar** - para cada posto de cobertura é calculado o posto permanente mais próximo baseado na distância euclidiana.

4 Aplicação

Esta seção detalha a implementação prática das etapas descritas no método.

A partir dos dados dos postos de monitoramento de tráfego do PNCT, foram integradas informações geográficas provenientes de diversas fontes com o objetivo de caracterizar cada posto de maneira consistente. São descritos, de forma sequencial, os procedimentos adotados para aquisição, tratamento e preparação dos dados, bem como a aplicação do algoritmo de clusterização *K-means*.

4.1 Aquisição e preparação de dados

As informações coletadas dos postos do DNIT são de identificação: código de identificação do posto [ID]⁵, estado em que se situa [uf], rodovia [br] e o quilômetro da rodovia [km]. Esses dados são extraídos da API do PNCT, isso é feito através do código em Anexo 1 para os postos permanentes e por Anexo 2 para os postos de cobertura. Os códigos são explicados por comentários em cada passo aplicado.

Adicionalmente às colunas já mencionadas, é criada a coluna [Tipo de Coleta], que contém a resposta ‘Permanente’ ou ‘Cobertura’ para cada posto. O resultado das duas requisições é mesclado em uma única tabela através do código em Anexo 3 à qual são adicionadas as demais colunas de caracterização geográfica.

Para a caracterização geográfica de cada posto, estão incluídas informações provenientes de diferentes fontes. A seguir, a lista de fontes e a informação que foi agregada a partir de cada uma.

a) Sistema Nacional de Viação (SNV):

- superfície da rodovia [sup], incluindo as respostas: ‘PAV’, ‘IMP’, ‘DUP’ e ‘PLA’, abreviações para pavimentada, implantada, duplicada e planejada, respectivamente;
- jurisdição [jur] com respostas: ‘Federal’ e ‘Estadual’;
- extensão [ext] da malha rodoviária total presente em um *buffer*⁶ projetado de 10 km ao redor de cada posto.

O conjunto de dados geoespaciais da malha rodoviária do Brasil⁷ é baixado, em formato *shapefile*, da pasta 202501A.

⁵Palavras em colchetes como [ID] e [uf] estão assim para simbolizar que são colunas da tabela de atributos dos postos, apresentada pela Figura 10.

⁶*buffer* é uma área ao redor de um objeto espacial. A medida em km associada representa o raio a partir de cada posto. Neste trabalho, buffers são usados em mais de uma variável e para cada uma é usado um tamanho escolhido com base na necessidade de área mínima para dar um resultado representativo.

⁷disponível em <[28](https://servicos.dnit.gov.br/dnitcloud/index.php/s/oTpPRmYs5AAAdiNr?path=%2FNSNV%20Bases%20Geom%C3%A9tricas%20(2013-Atual)%20(SHP)>”.></p></div><div data-bbox=)

O arquivo vetorial da malha rodoviária é lido e é realizada a junção entre os postos de contagem do PNCT e as informações da base do SNV, salvando o resultado como um novo arquivo *GeoPackage* denominado `postos_snv.gpkg`. Esse arquivo é incrementado com todas as demais colunas de caracterização dos postos. Os valores de jurisdição e superfície são atribuídos com base na linha (rodovia) mais próxima de cada posto, e a extensão da malha é calculada somando o comprimento total das vias dentro de um *buffer* de 10 km.

b) Instituto Brasileiro de Geografia e Estatística (IBGE):

- população [pop] calculada com base em um *buffer* de 50 km ao redor de cada posto, multiplicando a população dos municípios pela porcentagem da área destes que está dentro do *buffer*;
- renda per capita [rpc] calculada pela divisão da renda pela população, sendo a renda calculada da mesma maneira que a população.

Do IBGE é usado o censo demográfico de 2010 e a malha municipal em conjunto para associar o posicionamento geográfico aos dados censitários. A malha municipal usada é também de 2010 apesar de haver versões mais recentes disponíveis, isso é feito com o objetivo de garantir a compatibilidade temporal dos dados. A extração é feita pela API primeiramente da malha municipal e depois é adicionado a cada município os dados de população, renda e renda per capita.

O código de Python em Anexo 4 calcula a população e renda per capita para cada posto considerando um *buffer* de 50km. Adiciona essas duas informações ao arquivo `postos` que tem as colunas que caracterizam os postos, conforme listado e do SNV: superfície, jurisdição e extensão da malha ao redor dos postos.

c) *Shuttle Radar Topography Mission* (SRTM):

- [relevo] representa o desvio padrão da elevação em um *buffer* de 30 km.

É feito o download de arquivos ZIP com dados do SRTM de todos os estados brasileiros⁸, que são descompactados e é criado um mosaico único combinando as imagens de elevação. O mosaico é reprojetoado para o CRS EPSG:5880, utilizado no Brasil, e salvo como um arquivo raster que contém a elevação do terreno do Brasil.

É, então, calculado o desvio padrão das elevações do terreno a partir dos dados SRTM para cada posto de contagem de tráfego, usando um *buffer* de 30 km ao redor de cada posto. O

⁸disponível em <<https://www.cnpm.embrapa.br/projetos/relevobr/download>>

valor do desvio padrão, que reflete a variação da elevação na área, é adicionado ao arquivo de postos (Figura 10) na coluna `[relevo]`, e o arquivo `postos_snv.gpkg` é atualizado.

d) **OpenStreetMap (OSM):**

- número de pontos de interesse `[pois]` dentro de um *buffer* de 50 km.

Os dados de pontos de interesse (POIs) são baixados do *OpenStreetMap*⁹. É aplicado o filtro para as categorias `land_use` (`industrial`, `comercial`), `shop`, `amenity`, `leisure` e `tourism`, e é realizada a criação de *buffers* de 50 km ao redor dos postos de contagem. A partir desses *buffers*, é calculado o número de POIs presentes em cada área e essa contagem é adicionada à tabela de postos no arquivo `postos_snv.gpkg`.

e) **Agência Nacional de Águas (ANA):**

- distância dos postos até o porto público mais próximo `[dist_p]`;
- distância dos postos até o aeroporto público mais próximo `[dist_a]`.

O conjunto de dados geoespaciais de portos¹⁰ e de aeroportos¹¹ são arquivos *shapefile* com a localização dos portos e aeroportos no Brasil. É aplicado um filtro para selecionar apenas os portos e aeroportos públicos.

Para cada posto de tráfego, é calculada a distância euclidiana até cada um dos portos públicos. A menor distância, em quilômetros, é atribuída à coluna `dist_p` da camada de postos. O mesmo processo é feito para os aeroportos, com o valor sendo atribuído à coluna `dist_a`.

f) **Polícia Rodoviária Federal (PRF):**

- quantidade de acidentes `[acid]` que aconteceram dentro do *buffer* de 10 km.

A tabela utilizada é da PRF, dos acidentes de 2023 agrupados por ocorrência¹². É feita a conversão do formato CSV para GPKG, utilizando as colunas de latitude e longitude de cada ocorrência.

Os dados são, então, convertidos para o mesmo sistema de referência de coordenadas (CRS) EPSG:5880. Para cada posto, é criado um *buffer* de 10 km e, em seguida, calculado o número

⁹disponível em <<https://download.geofabrik.de/south-america/brazil-latest.osm.pbf>>.

¹⁰disponível em <<https://metadados.snirh.gov.br/geonetwork/srv/api/records/0afc9687-db93-4eb1-ab31-3bbd871ff303>>.

¹¹disponível em <<https://metadados.snirh.gov.br/geonetwork/srv/api/records/ce6bb0fe-ccfc-4cbf-9ff6-929ba80b47db>>

¹²disponível em <<https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/dados-abertos-da-prf>>.

de acidentes dentro desse *buffer*. A contagem é adicionada ao arquivo `postos_snv.gpkg`, na coluna "acid".

g) Secretaria Nacional de Trânsito (SENATRAN):

- [frota] de veículos calculada a partir da tabela de veículos por município e seguindo a lógica de ponderação da área de cada município dentro do *buffer* de 50 km ao redor de cada posto.

O SENATRAN disponibiliza dados da frota com algumas agregações. Para este trabalho, é usada a frota de dezembro de 2024 por município e tipo de veículo¹³. É baixado o arquivo CSV, que é tratado removendo o cabeçalho duplo do arquivo original e deixando-o em formato tabular para consumo. É usada a coluna `Total`, que soma todos os tipos de veículos de transporte rodoviário apurados por município. É lido: o arquivo processado de frota, a malha territorial dos municípios e os dados dos postos de tráfego. É feito um tratamento dos nomes dos municípios para padronização. São projetados *buffers* de 50 km ao redor de cada posto e é realizada uma interseção espacial entre esses *buffers* e os municípios. Para cada posto, calcula-se a frota proporcional com base na área do município que está dentro do *buffer*, e adiciona esse valor à coluna de frota no arquivo de postos.

h) Sistema IBGE de Recuperação Automática (SIDRA):

- produção agrícola total [prod] calculada através da ponderação da produção de cada município e de sua área dentro de um *buffer* de 50 km.

É utilizada a tabela 5457 de “Área plantada ou destinada à colheita, área colhida, quantidade produzida, rendimento médio e valor da produção das lavouras temporárias e permanentes”, a qual apresenta a produção através da variável definida como “Área plantada ou destinada à colheita (Hectares)”. A tabela, disponível em: <<https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/>>, é tratada para formato tabular. É feita a padronização dos nomes dos municípios e é criado o *buffer* de 50 km ao redor dos postos. É calculada a produção agrícola total de cada posto, ponderando a produção municipal e sua área dentro do *buffer* do posto. O valor da produção ponderada é adicionado à tabela de postos (Figura 10), e os dados são salvos no arquivo *GeoPackage*.

As variáveis utilizadas no presente estudo para o agrupamento dos postos de monitoramento refletem diferentes aspectos relevantes à caracterização da rodovia e à previsão da demanda de

¹³disponível em <<https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2024>>.

tráfego. Variáveis como população, renda per capita e frota são indicadores socioeconômicos que ajudam a estimar o volume potencial de viagens geradas ou atraídas na região, sendo, portanto, fundamentais para o entendimento da intensidade e regularidade da demanda. Já produção agrícola, distância a portos e distância a aeroportos relacionam-se diretamente à função logística das rodovias e são especialmente relevantes para identificar rotas com forte componente de transporte de carga e alta sazonalidade, como no escoamento de safras. A variável relevo, por sua vez, influencia o traçado e o nível de serviço das rodovias, podendo implicar em limitações operacionais, especialmente para veículos pesados. A extensão da malha rodoviária no entorno dos postos e a *presença de pontos de interesse* indicam o grau de conectividade e atratividade local, o que pode impactar diretamente o volume e a diversidade do tráfego. Por fim, jurisdição e tipo de superfície refletem aspectos institucionais e estruturais que influenciam tanto a priorização de investimentos quanto a qualidade da operação viária. Assim, a combinação dessas variáveis proporciona uma visão abrangente e multifacetada da funcionalidade e demanda das rodovias rurais, subsidiando análises mais robustas no contexto do planejamento de transportes.

A Figura 10 apresenta as cinco primeiras linhas da tabela de atributos dos postos de monitoramento de tráfego. E a Figura 11 mostra a dispersão geográfica dos postos e das rodovias utilizadas no cálculo da coluna de extensão.

Figura 10: Amostra da tabela de postos

fid	id	uf	br	km	Tipo de Coleta	sup	jur	ext	rpc	pop	relevo	pois	dist_p	dist_a	acid	frota	prod
1	76316531-cd90-463c-81bb-013b6f9345ad	PR	376	199,95	Permanente	DUP	Federal	38,209	18172,54	1009395	247,12	21	382,187	27,014	62	1047962	912256,65
2	e5f7ff1b-ea73-49de-a90c-02326614d0ac	MG	262	627,55	Permanente	PAV	Federal	21,578	4041,49	104176	427,76	2	476,187	47,663	8	85411	223466,3
3	e60d991e-3b0c-46c5-9aaf-0259de1069b1	RS	116	226,03	Permanente	PAV	Federal	30,798	35938,14	3027036	210,1	156	103,211	11,15	118	2324611	118481,71
4	00b7f256-25fa-4fd6-b9b0-026e9cf0ce8	PE	101	140,65	Permanente	DUP	Federal	25,021	6612,82	1423092	156,97	22	62,71	76,803	18	538101	288208
5	d032ffc8-694f-4649-bc4c-02bcd84c6292	PR	476	265,85	Permanente	PAV	Federal	22,534	4427,46	214516	336,6	0	170,836	26,668	28	175176	364311,49

Para preparar os dados para o agrupamento dos postos, é necessário transformar colunas categóricas em numéricas. O processo realizado inclui o *one-hot encoding* para a coluna *sup* e a codificação binária para as colunas *jur* e *Tipo de Coleta*, por essas conterem apenas duas categorias possíveis. O código de *Python* em Anexo 5 realiza esse processo, e os passos são explicados por comentários no código.

A tabela de saída do código em Anexo 5 não apresenta as colunas *Tipo de Coleta*, *sup* e *jur*, mas sim uma coluna para cada valor possível de *sup*, e uma única coluna binária para *Tipo de Coleta* e *jur*. A Tabela 1 apresenta os valores dessas colunas na forma categórica. A Tabela 2 mostra os valores convertidos para a representação numérica binária.

Figura 11: Postos de monitoramento de tráfego e rodovias

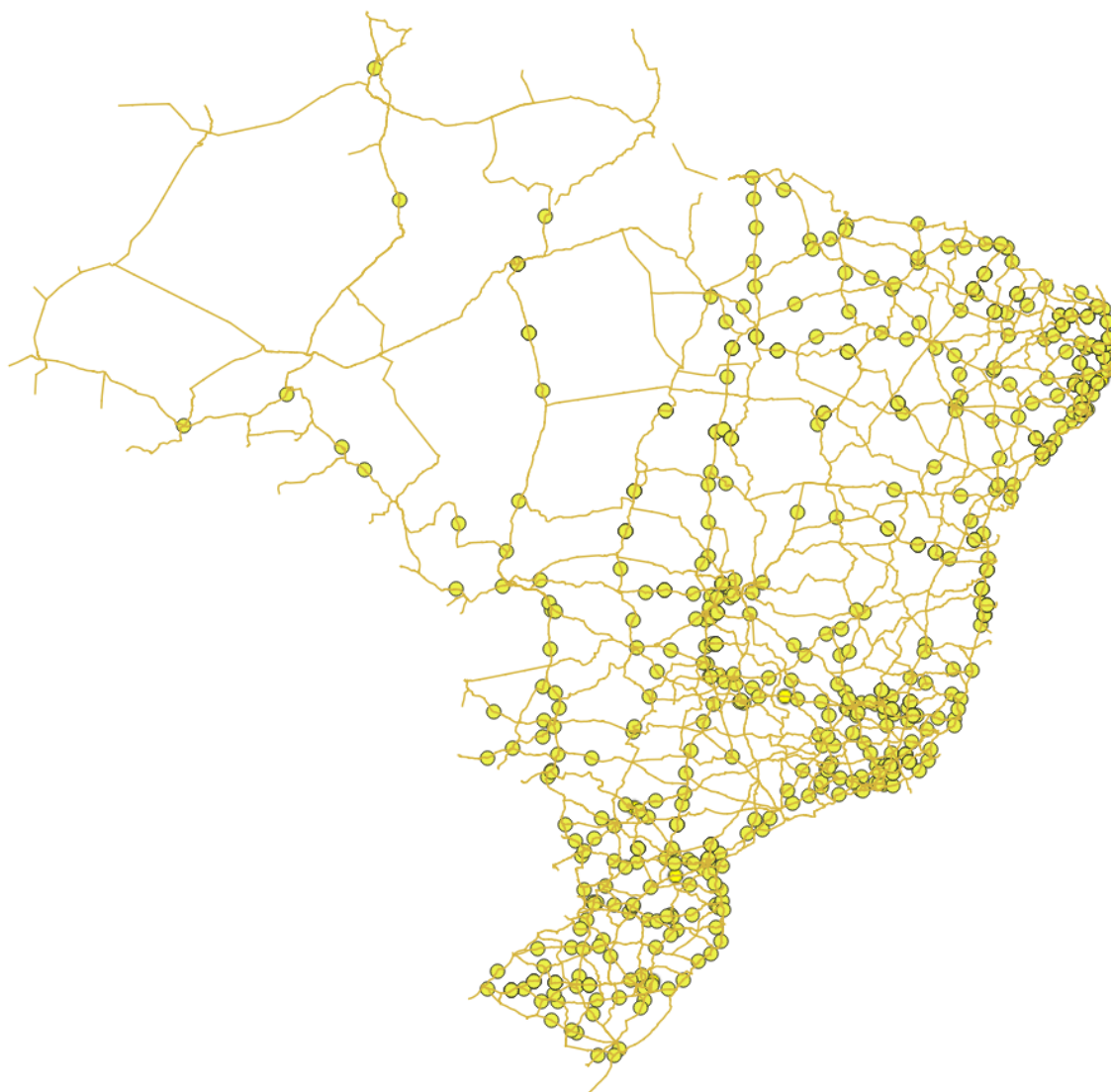


Tabela 1: Atributos categóricos de um posto exemplo

id	Tipo de Coleta	sup	jur
76316531-cd90-463c-81bb-013b6f9345ad	Permanente	DUP	Federal

Tabela 2: Codificação binária dos atributos categóricos

id	Tipo_de_Coleta_Permanente	jur_Federal	sup_DUP	sup_IMP	sup_PAV	sup_PLA
76316531-cd90-463c-81bb-013b6f9345ad	1	1	1	0	0	0

4.2 Agrupamento dos postos

A clusterização dos postos de monitoramento de tráfego é feita através do algoritmo *K-means* desenvolvido pelo código de *Python* em Anexo 7. O código utiliza a biblioteca *scikit-learn* para realizar o agrupamento dos postos com base em variáveis específicas.

É feita a leitura da tabela de postos com atributos categóricos codificados e a normalização dos dados é feita utilizando o *StandardScaler* para garantir que todas as variáveis tenham a mesma escala. O *StandardScaler*, para cada variável, subtrai a média e divide pelo desvio padrão. Assim, cada coluna (variável) passa a ter média igual a 0 e desvio padrão igual a 1, diferentemente da padronização via *MinMaxScaler*, que coloca os valores dentro do intervalo $[0, 1]$, *StandardScaler* não impõe limite para seus valores. O valor de cada ponto em cada variável representa a quantos desvios padrão ele está da média (valores positivos indicam acima da média e negativos, abaixo).

Para a clusterização são utilizadas todas as colunas de caráter descritivo do posto presentes na Figura 10 exceto Tipo de Coleta, e não são usadas as de identificação (id, uf, br, km).

O código faz o cálculo do Índice de Silhueta para identificar o número ideal de *clusters*, variando de 2 a 10, e escolhe o número ótimo com base no maior *Silhouette Score*. Paralelo ao Índice de Silhueta, é desenvolvido outro código, em Anexo 6 que aplica o método do cotovelo. Neste código é feita a plotagem do somatório de erros quadráticos para cada número de *clusters* no intervalo de 2 a 10 e analisando a curva, é possível determinar o número ideal de agrupamentos a ser usado.

Ambos os métodos, Índice de Silhueta e Cotovelo, são explicados na Seção 2.4.1. O algoritmo *K-Means* é então aplicado usando o número ótimo de *clusters* considerando os dois métodos de sua determinação. A escolha final é explicada na Seção 5 de resultados.

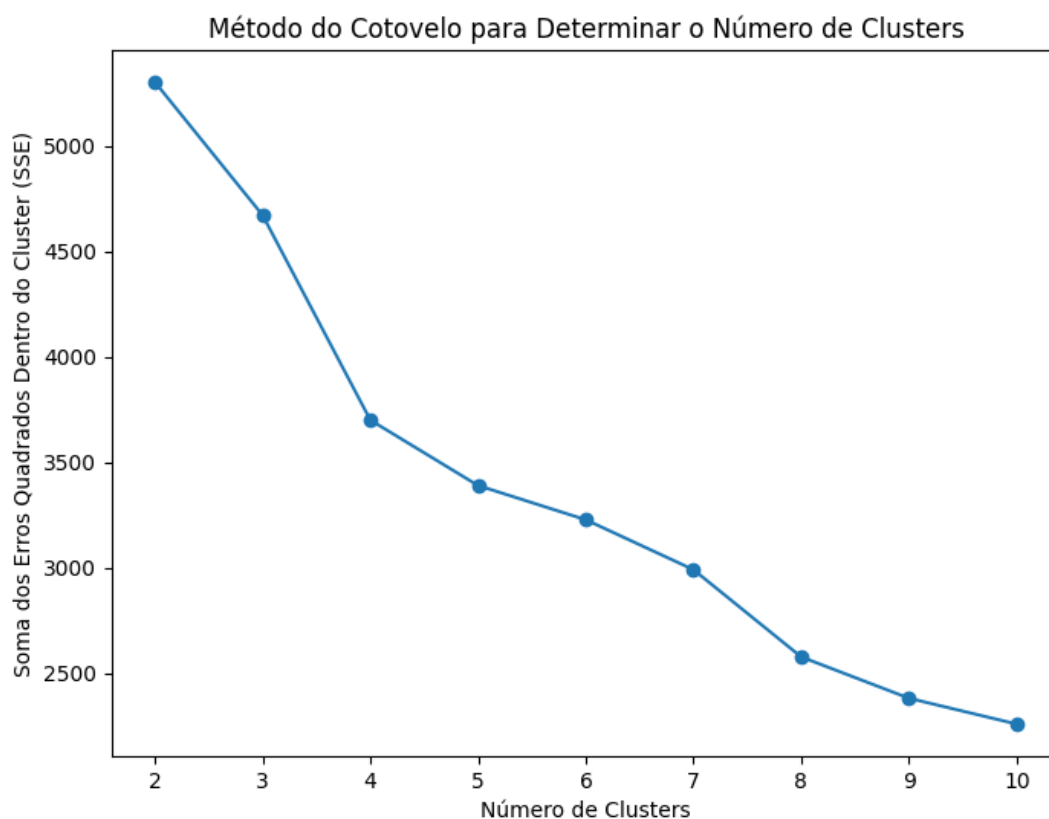
Os centroides dos *clusters*, a quantidade de postos atribuídos aos *clusters* e importância de cada variável nesta atribuição são resultados gerados pelo código em Anexo 7 e também são apresentados na Seção 5.

5 Resultados

Esta seção apresenta os resultados da aplicação do algoritmo de clusterização, iniciando pela avaliação do número ideal de *clusters* com base no método do cotovelo e no índice de silhueta. Em seguida, são apresentadas as visualizações dos grupos formados, acompanhadas de interpretações sobre suas principais características. Por fim é comentado sobre a associação entre os postos de cobertura e os postos de coleta contínua mais semelhantes, e é apresentada a Tabela 5 resultante.

A execução de Anexo 6 gera a Figura 12 com o resultado do método do cotovelo que fornece a primeira indicação quanto ao número ideal de *clusters*.

Figura 12: Cotovelo

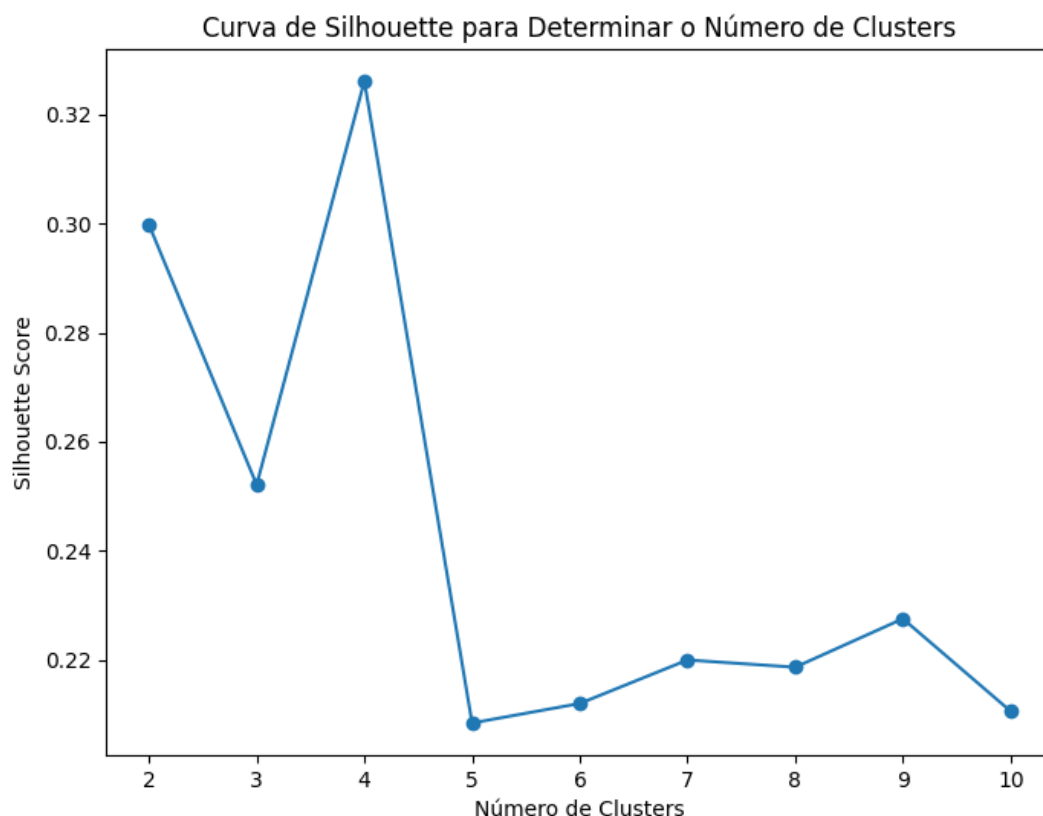


O ponto em que a curva diminui sua inclinação significativamente é com 4 *clusters*.

O código em Anexo 7 calcula o *silhouette score* para cada número de *clusters*, apresentado na Figura 13.

No método de silhueta, o maior *silhouette score* indica o número ideal de *clusters*. Portanto, os dois métodos coincidem em 4 como o número de *clusters* a ser usado.

Figura 13: Índice de silhueta



O algoritmo em Anexo 7 é executado com a variável `optimal_n_clusters = 4`. E a Figura 14 apresenta, em ordem, as variáveis mais importantes na determinação do *cluster* de cada posto. Isso é determinado somando, para cada variável, os valores absolutos das coordenadas dos centroides naquela variável (naquele eixo). Isso mede quanto os centroides variam entre si naquela dimensão e quanto maior a variação, mais aquela variável está influenciando na separação dos grupos.

Com base na Figura 14, é possível observar que as variáveis mais influentes na formação dos clusters são `sup_PAV`, `sup_PLA`, `jur_Federal` e `sup_DUP`, todas relacionadas à infraestrutura e jurisdição da via. Já variáveis como `frota`, `pop` e `acid` também se destacam, indicando que aspectos operacionais e demográficos contribuem significativamente para a diferenciação entre os grupos. Em contrapartida, variáveis como `dist_a`, `dist_p` e especialmente `sup_IMP` apresentam baixa influência. No caso de `sup_IMP`, esse resultado é justificado pelo fato de que apenas 1 dos 430 postos analisados apresenta valor igual a 1, sendo todos os demais iguais a 0, o que reduz drasticamente sua capacidade de contribuir para a separação dos grupos.

A quantidade de postos pertencentes a cada *cluster* está apresentada na Tabela 3.

Figura 14: Importância das variáveis no agrupamento

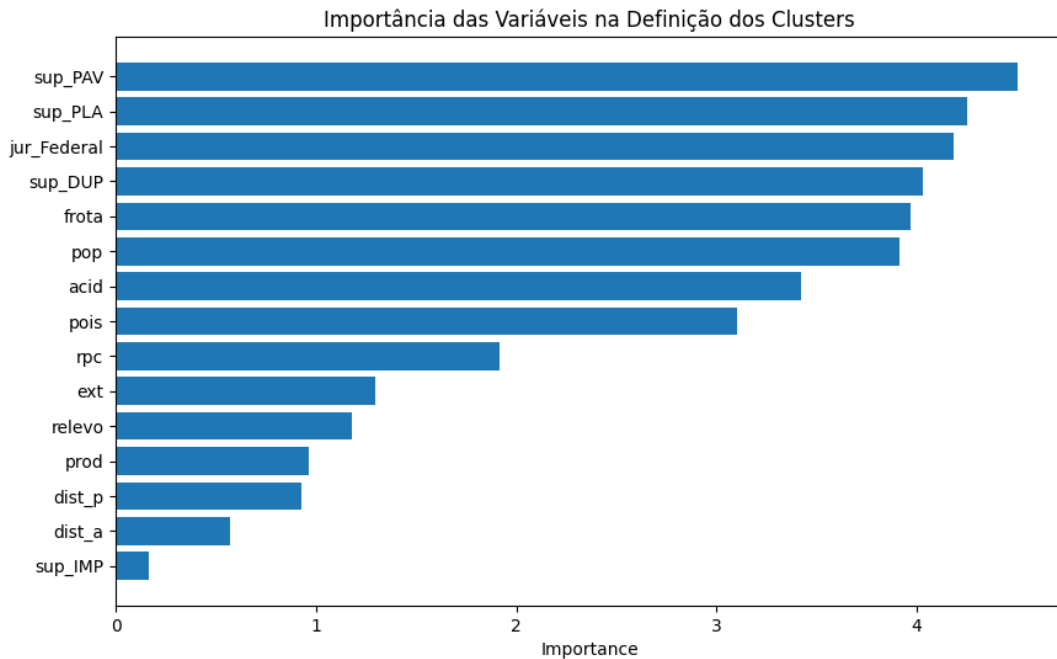


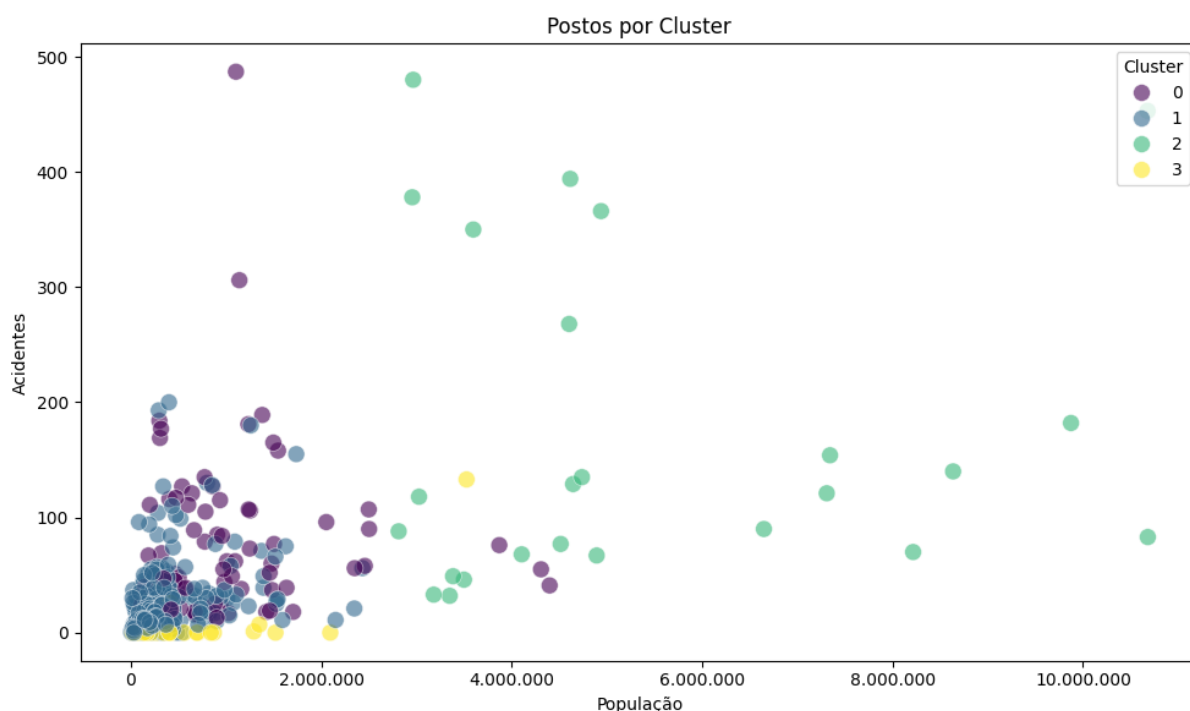
Tabela 3: Quantidade de postos pertencentes a cada *cluster*

<i>Cluster</i>	Quantidade de postos
0	74
1	296
2	25
3	35

A Figura 15 traz uma primeira visualização do agrupamento e alguns padrões podem ser notados quanto ao impacto das variáveis de população e quantidade de acidentes no entorno dos postos. Nesse caso as duas variáveis tem correlação já que quanto mais pessoas, em geral, mais veículos e consequentemente mais possibilidade de acidentes.

É possível notar alguns padrões nos *clusters* analisando a Figura 15. Os postos do *cluster 2* apresentam os maiores valores de população, tendo exemplos com baixa e com alta taxa de acidentes. O *cluster 3* têm visivelmente baixa taxa de acidentes em comparação com a população (pontos na parte baixa do gráfico). O inverso é observado para os postos do *cluster 1* que tem maiores valores de acidentes mesmo ocupando a parte esquerda do gráfico (baixa população relativamente). Os postos do *cluster 0* apresentam, assim como os do *cluster 2*, um equilíbrio (gráfico) das duas variáveis, contudo se concentram mais perto da origem (mais baixos valores dos dois eixos em exibição) em comparação com *cluster 2*.

Figura 15: Dispersão dos postos por *clusters* - eixos: população × acidentes



Os centroides de cada posto são marcados com X e um número relativo ao *cluster* na Figura 16, nela os valores dos eixos são dos dados normalizados para poder adequar o posicionamento dos centroides, que são calculados nessa escala.

Foi colocado na camada geoespacial de postos o *cluster* de cada um e a Figura 17 apresenta essa distribuição.

A partir da visualização geográfica dos postos também é possível notar alguns padrões nos *clusters*. O grupo 1, apesar de haver alguns postos perto do oceano, em sua maioria estão no interior do país, que em geral tem baixa concentração populacional, corroborando com a análise da Figura 16. Os postos do *cluster* 2 são localizados, quase exclusivamente em cidades grandes, como: Belo Horizonte, Rio de Janeiro, Fortaleza, Curitiba e Porto Alegre. Isso indica que há forte relação desse grupo de postos com alta população em seu entorno.

Para checar as hipóteses levantadas ao analisar os resultados é feito o cálculo, pelo código em Anexo 8, da média de cada variável usada no agrupamento dos postos. Os valores estão apresentados na Tabela 4.

A Tabela 4 apresenta as médias das variáveis utilizadas no agrupamento para cada um dos quatro *clusters*. O significado de cada variável é explicado na Seção 4.1. A interpretação dos resultados permite inferir características distintas entre os grupos formados.

O *cluster* 0 agrupa postos localizados em regiões com população relativamente alta (1.044.192,20 habitantes) e renda *per capita* expressiva (R\$ 8.792,65), além de valores media-

Figura 16: Dispersão dos postos e centroides por *clusters* - eixos: população × acidentes

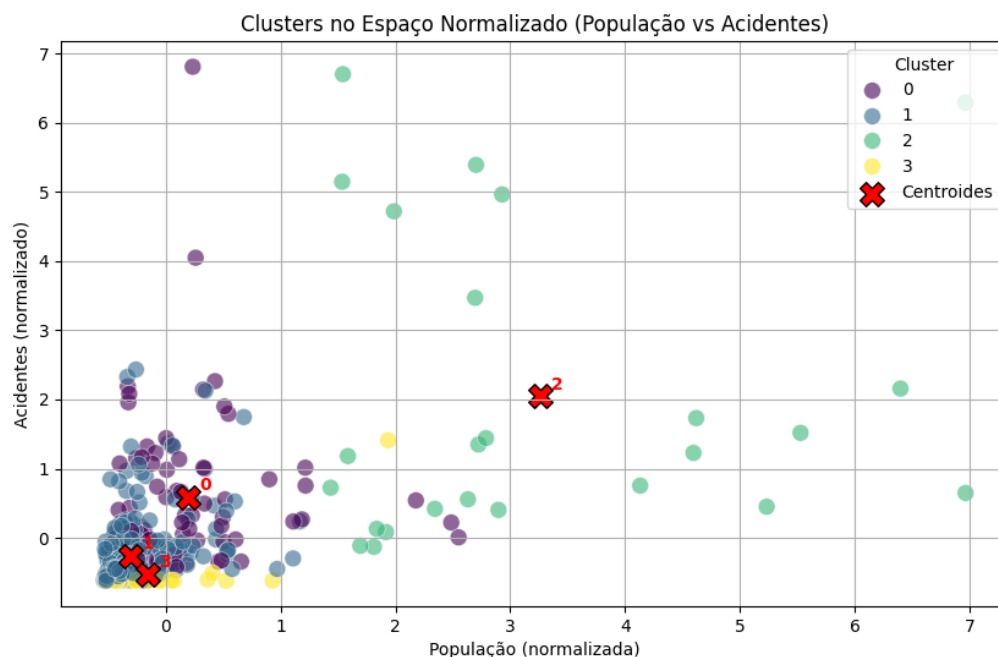
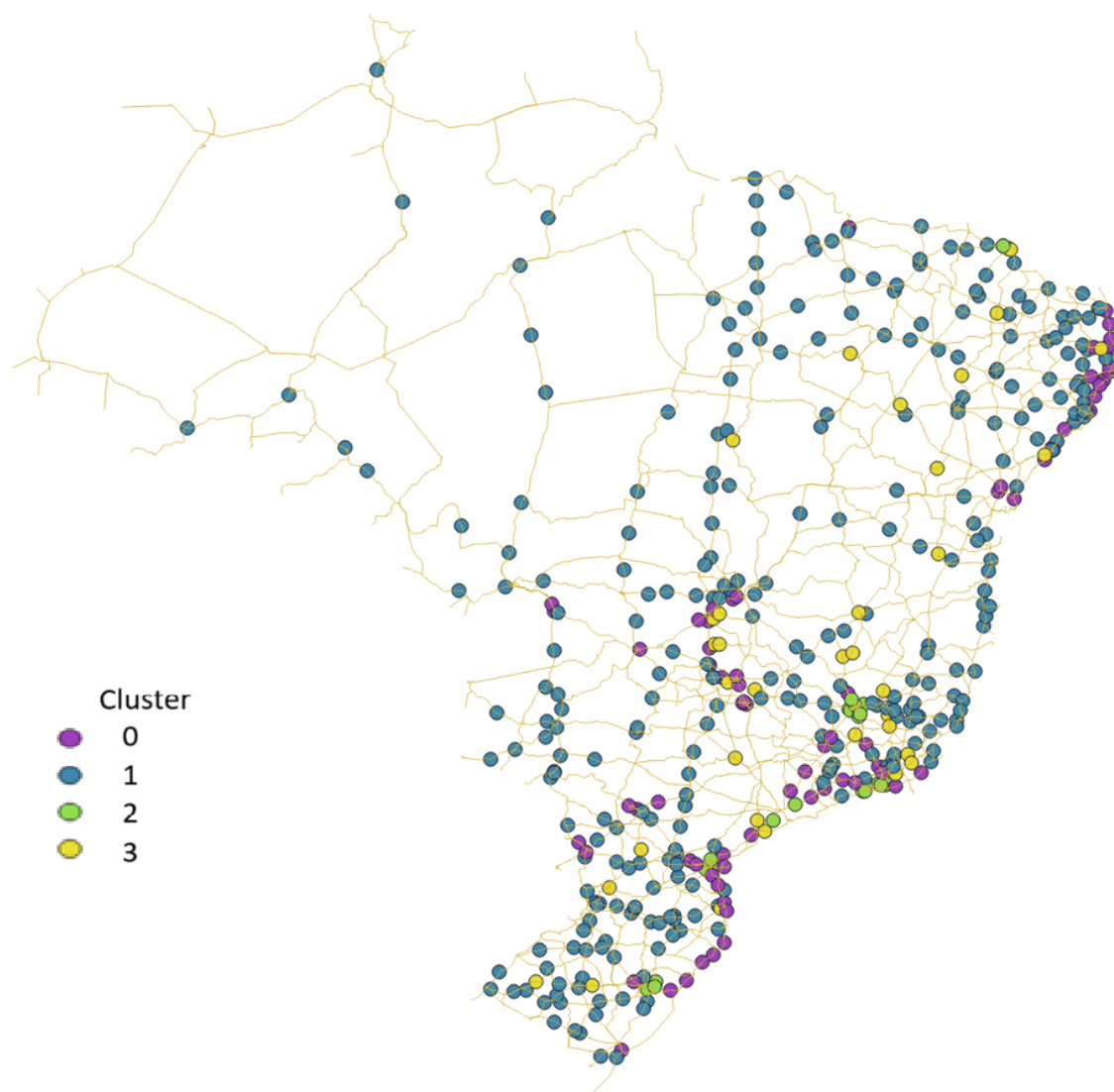


Tabela 4: Médias das variáveis para cada *cluster* de postos

cluster	0	1	2	3
extensão da malha [km]	32,894	28,092	35,372	17,762
renda <i>per capita</i>	R\$ 8.792,65	R\$ 5.360,04	R\$ 14.985,74	R\$ 6.369,58
população	1.044.192,20	339.150,27	5.408.351,28	547.625,43
relevo (desvio padrão da elevação)	221,89	207,64	308,66	272,05
pontos de interesse	21,88	10,59	200,92	30,31
distância ao porto público mais próximo [km]	223,19	341,03	124,28	304,04
distância ao aeroporto público mais próximo [km]	29,19	36,14	30,28	40,31
acidentes	79,19	23,52	174,84	6,06
frota	744.360,62	257.021,78	3.737.462,76	413.717,26
produção agrícola	226.198,01	185.018,47	61.038,07	155.911,56
jur_Federal	1,00	1,00	1,00	0,03
sup_DUP	1,00	0,00	0,64	0,00
sup_IMP	0,00	0,003	0,00	0,00
sup_PAV	0,00	0,997	0,36	0,00
sup_PLA	0,00	0,00	0,00	1,00

Figura 17: Distribuição espacial dos postos, coloridos por *cluster*



nos de frota, produção agrícola e acidentes. Esses postos estão em trechos totalmente federais e com superfície predominantemente duplicada. Os pontos de interesse e o relevo também apresentam valores intermediários. **Perfil:** regiões economicamente relevantes, com infraestrutura consolidada e tráfego moderado.

O *cluster* 1 reúne os postos com os menores valores médios em diversas variáveis: é o grupo com menor população, renda *per capita*, frota, pontos de interesse e variabilidade de relevo. A jurisdição é federal, e a maioria dos trechos é pavimentada. Apresenta pequena quantidade de acidentes e relativamente alta produção agrícola. **Perfil:** áreas de infraestrutura básica, baixa densidade populacional e econômica, representando regiões menos dinâmicas do ponto de vista do tráfego.

O *cluster* 2 apresenta os maiores valores médios em diversas variáveis: população (5,4 milhões), frota (3,7 milhões), acidentes (174,84), renda *per capita*, extensão da malha e pontos de interesse. A jurisdição da rodovia é federal em todos os postos, com predomínio de superfície pavimentada. Por outro lado, é o grupo com a menor média de produção agrícola (61.038,07). **Perfil:** áreas urbanas, com intensa atividade econômica, grande população e quantidade de veículos e complexidade viária elevada.

O *cluster* 3 compreende postos em regiões com valores intermediários de população, frota e produção agrícola, mas apresenta uma particularidade: todos os trechos estão em superfície planejada e a jurisdição é majoritariamente estadual. Tem alta variabilidade de relevo e a menor média de acidentes entre todos os grupos (6,06). **Perfil:** regiões com malha rodoviária em fase de planejamento ou expansão, inseridas em contextos estaduais e com menor maturidade em termos de infraestrutura.

Apesar da utilidade da análise, é importante lembrar que, conforme apontado por Schneider e Tsapakis (2009), uma das principais desvantagens dos métodos baseados em *cluster analysis* é que a lógica por trás da formação dos grupos pode ser de difícil interpretação. Isso se torna ainda mais relevante quando os grupos formados não correspondem a categorias intuitivamente reconhecíveis ou operacionais, exigindo análise crítica e conhecimento de domínio para validar os resultados.

O passo sucessor das análises por *cluster* é o estabelecimento de pares de postos. É calculado, para cada posto de cobertura, o posto permanente mais próximo no espaço vetorial das mesmas variáveis usadas no agrupamento. Ser o mais próximo nesse espaço vetorial significa ser o mais similar segundo as variáveis (ou atributos) utilizadas. O código em Anexo 9 realiza os seguintes passos: as variáveis são normalizadas usando *StandardScaler* e transformadas em vetores numéricos. Em seguida, os postos são separados conforme o tipo de coleta e organizados em duas matrizes, uma para cobertura e outra para permanentes. A função *cdist* é utilizada para calcular todas as distâncias euclidianas entre os vetores, e o par com menor distância é selecionado para cada posto de cobertura.

Os pares resultantes são exibidos pela Tabela 5. Os postos estão identificados pelo seu ID seguido da UF e número da rodovia. A terceira coluna é a distância euclidiana entre os dois postos no espaço vetorial das variáveis normalizadas por *StandardScaler*.

Tabela 5: Postos de cobertura e o posto permanente mais similar

Posto de Cobertura	Posto Permanente mais Similar	Distância Euclidiana
096760ba-e9f2-4304-863d-007101ddcee4 (UF: PA, BR: 163)	e518daaa-1c6d-4f8e-a74f-5d8d06ca555e (UF: PA, BR: 163)	2,351153
10d8466e-15cc-4577-b8e3-03b4f62c4810 (UF: RS, BR: 287)	af9be5c4-0d9d-47c4-be58-fd619a96a60d (UF: SP, BR: 364)	2,213512
96bda42b-7b02-4b38-ae67-09fd38e350e6 (UF: RN, BR: 226)	1b7c3bcc-335d-4ca3-86d7-3bb83eee0783 (UF: PB, BR: 230)	0,602973
aefa9552-2676-4a92-9da1-0d7456ba0f81 (UF: PR, BR: 476)	c0f513bb-3e29-4ee3-979d-1af641df1bf3 (UF: MG, BR: 381)	2,51431
c850bcbf-30a9-459c-a7b2-14c7f9e9f6ed (UF: MG, BR: 262)	ef5e143e-8121-4810-b43d-bfa8a3ad4b53 (UF: MG, BR: 259)	1,004597
c132428c-acdf-4fd9-bcff-e15f470d6fc5 (UF: TO, BR: 153)	50a71689-06cd-4c26-9a39-19729becb4c1 (UF: RO, BR: 364)	0,592022
ffe18d83-3166-40c8-b5e5-1ea4015d112e (UF: CE, BR: 122)	ce40690d-cd1e-434a-9947-15ff319e9a26 (UF: PE, BR: 122)	1,252612
5fb7efde-e2cb-4e7f-a2bd-2d7fb0f737f (UF: MA, BR: 10)	1dc8905f-cf0c-48f4-9949-04e718972a46 (UF: PE, BR: 122)	1,023408
2fc8c3fb-e5f7-4e25-8e12-2d855e477740 (UF: AL, BR: 101)	c5dea4a7-0c9f-44f2-ac5d-a6898916e4ec (UF: SE, BR: 235)	1,811581
895a4395-d451-4c06-8b18-312eb6ea1b19 (UF: BA, BR: 430)	261d24d-980f-4ee4-b38f-5ad6ee180409 (UF: MG, BR: 365)	0,381948
8a15431a-bfb8-4798-afbb-33b24af7e87e (UF: PE, BR: 101)	00b7f256-25fa-4fd6-b9b0-026e9cf0ece8 (UF: PE, BR: 101)	0,720326
8e5afb89-1b94-4b7f-bef9-33c17d8119b2 (UF: GO, BR: 457)	94e830ab-289f-42a9-9a4b-9989c35dd8c5 (UF: GO, BR: 352)	1,031794
ccdedb51-8ffc-48dd-b3e1-369aa80c034c (UF: PE, BR: 423)	49aao670-0517-4436-929e-79c58440c49a (UF: PE, BR: 423)	0,260945
276cdbd4-9510-4595-80b9-3a6332587d37 (UF: BA, BR: 101)	9d5e37b2-eb15-4f2f-8841-ccfd86e6f7af (UF: RS, BR: 471)	0,814758
8f66578c-88e1-410e-99f5-3ed2b339feec (UF: GO, BR: 414)	0be9f75a-8cc7-4d1f-8d2c-83f2ce8b585a (UF: GO, BR: 158)	1,143619
ea3317fe-bbe9-4090-bf1a-44a50d50d50a (UF: MG, BR: 474)	bce346e0-c3b1-424f-9693-2c28a05fdb9a (UF: MG, BR: 494)	0,718679
5c5b3120-5532-4d25-b865-459d48821537 (UF: RJ, BR: 393)	574a2cc1-31ea-4402-8299-0c26b0214002 (UF: SP, BR: 101)	1,374352
bea17a13-003a-45f9-b112-46656d1fd065 (UF: CE, BR: 20)	31671e51-a3a7-4bfe-83a7-e34784bbaa81 (UF: AL, BR: 423)	0,459129
1016aaea-f9a3-4b7b-9ae0-48ba03fb4c5b (UF: TO, BR: 80)	90ec25b8-3ae2-48f1-955b-6e7869b58108 (UF: MS, BR: 267)	0,939489
28682f8c-4bd4-43b7-b838-4deafbeca13c (UF: PE, BR: 104)	551fdb25-bbaa-490f-801b-2262f6afbda4 (UF: AL, BR: 104)	0,861584
33934b0b-8fd2-46e1-a219-5057f185c397 (UF: PA, BR: 230)	ad9181f7-01cf-4cca-b40b-d1576b3f81ab (UF: RJ, BR: 101)	0,867358
7df96d8a-5827-42da-a3de-51d0b6983c4b (UF: SE, BR: 235)	8d76d2f2-8b23-447f-8491-d511cbccdd86 (UF: BA, BR: 101)	0,793518
693284e0-3284-4c2f-942a-5a1953d23ce9 (UF: RS, BR: 290)	18fb7d24-d42d-4f0e-8044-edc63a6f4632 (UF: RS, BR: 290)	0,474073
a9bb37a2-3af5-4269-815c-5d1e8f19733c (UF: RS, BR: 377)	ce40690d-cd1e-434a-9947-15ff319e9a26 (UF: PE, BR: 122)	1,258225
0edc1990-9966-4275-89d2-642cb8b3ce96 (UF: MG, BR: 120)	574a2cc1-31ea-4402-8299-0c26b0214002 (UF: SP, BR: 101)	1,153144
5cbf9768-6788-476c-8474-64873d6ac634 (UF: BA, BR: 101)	cfe8c2ee-3e8b-4f74-b079-59434bb1ed9a (UF: BA, BR: 101)	0,79996
d4d2b16a-1a60-4eba-8aa3-65ff7089b042 (UF: PB, BR: 230)	cc4b15ac-2a54-47bb-b9f4-58b09aea6dac (UF: SE, BR: 101)	1,208057
53dad5b2-f480-44f4-9266-67c1da86816a (UF: RN, BR: 101)	00b7f256-25fa-4fd6-b9b0-026e9cf0ece8 (UF: PE, BR: 101)	1,404363
5a8178a9-ebd8-4df2-b7a9-6f5a16b385d1 (UF: PE, BR: 104)	22219ce1-c929-4a4b-8f1e-6fe9fb053013 (UF: CE, BR: 222)	0,784896
aecb5051-a58d-445a-b71e-7cd04b5c7a11 (UF: BA, BR: 367)	ae6b1442-f49c-4a10-8b5c-b0ec7b496aee (UF: MA, BR: 222)	0,517352
bfa15301-30f6-4498-a406-7d539dd9b629 (UF: RN, BR: 226)	4e115063-4716-4f30-a40d-fd5d2c86968b (UF: PB, BR: 230)	1,076444
e843579a-98d2-473d-bd84-8114df9de6df (UF: TO, BR: 10)	8b6fffb1-fc76-47c9-a30c-ff3d5a8fb62a (UF: GO, BR: 490)	1,56933
0105e65f-3889-4630-b9bd-825813f330b6 (UF: RJ, BR: 492)	94b239ec-249f-44d4-8be9-907da26a1dc5 (UF: SP, BR: 478)	0,938853
633352fc-6805-4a72-b4bb-8b4629d5eecd (UF: RN, BR: 427)	2dcba5d4-e759-4a4f-a893-b1da00f6f53e (UF: MG, BR: 116)	1,012568
8fe52aae-c3da-4dca-b2b1-8d520eda7563 (UF: RS, BR: 287)	6ab3da84-073c-bf7e-86c4-3259e4fdf1df (UF: RS, BR: 293)	0,697253
ec862dfd-b37a-4682-b7d9-93e73fce0fdd (UF: PI, BR: 324)	ce40690d-cd1e-434a-9947-15ff319e9a26 (UF: PE, BR: 122)	1,003115
7e46bc0e-3d95-44e2-b845-9583140f50a9 (UF: CE, BR: 20)	c07d871a-00ad-435d-8bd4-4c2106df11f9 (UF: BA, BR: 420)	0,82897
435bdd42-b48b-4e76-a5f3-982e4c404584 (UF: PR, BR: 476)	c07d871a-00ad-435d-8bd4-4c2106df11f9 (UF: BA, BR: 420)	1,14832
2b0e6492-63eb-4eb3-94d5-9b7bac30fc47 (UF: BA, BR: 407)	b7b4d26f-cd7d-417d-9236-3db7559fc38e (UF: MG, BR: 496)	1,098542
679b884d-43db-4250-abf4-9f0c6732d56a (UF: MT, BR: 158)	f0ea0a51-b9d8-44a2-bba4-d0451f6626de (UF: MT, BR: 364)	0,666873
7f885007-e437-4fd4-bf45-9f305866fa42 (UF: PR, BR: 466)	d260d49b-5922-42a3-a32b-2a09fd54d97f (UF: MG, BR: 452)	1,043522
a43d77dc-2965-495a-9280-a0765c8a2ea2 (UF: MT, BR: 158)	56957ff9-ebf3-4fa2-9c2d-f358113bdfb6 (UF: MT, BR: 364)	1,927214
e8b720a2-d15e-4cbf-b171-a728eaf57007 (UF: PE, BR: 408)	d8883184-5427-4bba-bee5-afc11041ee83 (UF: AL, BR: 316)	1,411204
1b35b32c-20c8-473e-852e-acc02691eec6 (UF: GO, BR: 490)	19421435-ac50-4fb2-9f8e-98528f5b40f1 (UF: MG, BR: 497)	1,091419
63207813-f34b-4ef5-8f57-aef3a09fc0ef (UF: MG, BR: 262)	5a6f3825-29ec-4500-85ee-f7f295776bc9 (UF: MG, BR: 262)	1,42002
54ab2012-98ad-41dd-b847-b11a241651fc (UF: MG, BR: 262)	efd016b0-b411-4350-80cd-e8435085f954 (UF: MG, BR: 262)	0,422003
0db7e712-7dbb-4767-9e77-b395f5b8db10 (UF: RJ, BR: 492)	94b239ec-249f-44d4-8be9-907da26a1dc5 (UF: SP, BR: 478)	1,412789
41d276bf-891d-4a07-aa48-b522c501c01d (UF: PI, BR: 324)	ce40690d-cd1e-434a-9947-15ff319e9a26 (UF: PE, BR: 122)	0,509869
7e28937c-5f90-4a4a-9a2b-b89d2e0c46f3 (UF: PB, BR: 230)	8140b571-f93b-458a-817e-e109cc7968a9 (UF: PB, BR: 101)	0,608939
72bbb685-530d-45fb-9ec6-bbb46e9d01cc (UF: MA, BR: 230)	9a6b06c6-4b6a-46b8-9ef1-41bba325e8f0 (UF: MS, BR: 158)	0,395096
1058953c-03a1-4460-8d9c-bc3758546dde (UF: PI, BR: 135)	f6b46e9f-c4f9-4d8a-875c-7edfc1bbef92 (UF: BA, BR: 135)	0,795384
aeaf6e9c-b4cf-4612-8124-c09a98db73d1 (UF: TO, BR: 153)	9a6b06c6-4b6a-46b8-9ef1-41bba325e8f0 (UF: MS, BR: 158)	0,499598
5185c4af-5075-4646-9058-c703feba73d5 (UF: BA, BR: 101)	31671e51-a3a7-4bfe-83a7-e34784bbaa81 (UF: AL, BR: 423)	0,621929
4555bf43-6659-4cea-9e60-cf256cf2bd75 (UF: CE, BR: 116)	11868f1d-6587-4446-b706-374604d91bb8 (UF: MA, BR: 316)	0,4987
e3b6ef91-1850-4a17-a0d4-cfbd7f448266 (UF: RJ, BR: 116)	4a890137-795f-4032-84da-9ea6681c01ce (UF: RJ, BR: 393)	0,548507
ca07e651-60e7-4672-a6a3-d20445a267d7 (UF: RN, BR: 226)	1b7c3bcc-335d-4ca3-86d7-3bb83eee0783 (UF: PB, BR: 230)	0,774232
bb2dc4a2-a5f3-4cb4-b3b7-bf1d89f10de (UF: GO, BR: 70)	c7e84d29-273a-4b37-a604-7a347265085b (UF: TO, BR: 242)	1,18942
b0bb94fa-7cf4-4e74-9f12-dc1e4c50256f (UF: AL, BR: 316)	fe43a3a9-ee76-48e7-8665-1d1b3088e976 (UF: RN, BR: 304)	1,23709
46c8e78b-cc36-40f0-a84e-de30e4460f23 (UF: PB, BR: 101)	8140b571-f93b-458a-817e-e109cc7968a9 (UF: PB, BR: 101)	0,693129
b8bcfce7-63a3-4890-9046-e03adb9b151b (UF: PA, BR: 158)	68b18009-364f-4476-9203-78a3bdaaf697 (UF: RS, BR: 392)	0,562023
6b8c0f3c-a8ba-4ae3-b907-e7de498aec01 (UF: BA, BR: 330)	9f2a25b2-95e0-4152-9703-67954f3118cf (UF: MA, BR: 135)	0,938019
eebeee8b-219a-4d03-a08c-e957518ac6b8 (UF: SE, BR: 349)	d479bb6b-c3d8-4072-a27b-04d416cd3664 (UF: MG, BR: 120)	2,20427
e0a60b14-0a1f-49f7-96a3-ee9095bf1271 (UF: PR, BR: 153)	a92410f9-a82d-4e33-ba84-a5561ec426a3 (UF: SC, BR: 282)	0,518515
7ce236a9-0644-42e8-951c-f4884c91a0d2 (UF: BA, BR: 30)	40a41f05-7f04-4e84-8305-ff76069ec6d6 (UF: BR: 423)	0,696709
abd7fa95-66b7-4485-aa76-ff94d7466c10 (UF: PB, BR: 408)	574a2cc1-31ea-4402-8299-0c26b0214002 (UF: SP, BR: 101)	4,106553

6 Conclusões

Os resultados obtidos neste estudo permitem concluir que a utilização de dados geográficos é suficiente para identificar padrões de características entre postos de monitoramento de tráfego. A estrutura dos *clusters* formados revelou coerência com características como jurisdição e infraestrutura da via, indicando que essas variáveis são decisivas no agrupamento dos postos. Mais do que agrupar dados semelhantes, a aplicação do algoritmo permitiu a construção de perfis interpretáveis para cada grupo, o que potencializa seu uso em estratégias operacionais e também de projeto e planejamento.

Durante o desenvolvimento do trabalho, constatou-se que há uma ampla disponibilidade de dados públicos gratuitos e relevantes para estudos dessa natureza, embora a padronização e integração entre as fontes demandem esforço técnico considerável. O processo de tratamento e limpeza de dados é necessário para lidar com exceções e garantir consistência temporal e espacial, evidenciando a importância de habilidades em programação para trabalhos desse tipo.

O uso de bibliotecas amplamente disponíveis em *Python*: *numpy*, *pandas*, *geopandas* e *scikit-learn*, demonstrou ser suficiente para o porte do estudo, o que reforça a viabilidade de replicação da abordagem por profissionais com conhecimento intermediário em ciência de dados. Como benefício, verificou-se que o método pode apoiar estratégias de expansão volumétrica por similaridade, atribuindo a postos de cobertura estimativas mais robustas a partir de padrões de postos permanentes com características semelhantes.

Assim, este trabalho estabelece uma base promissora, mas que demanda aprofundamento técnico para ser plenamente aproveitada como ferramenta de apoio à expansão volumétrica e à gestão inteligente de infraestrutura viária. Os *clusters* formados, podem servir para cálculo do fator de expansão médio entre os postos contínuos e criar um fator de grupo a ser usado nos postos de cobertura do respectivo *cluster*. E a Tabela 5 pode ser usada para aplicar os fatores de expansão do posto permanente no posto de cobertura da mesma linha. É importante testar os resultados de estimativas e obter um parâmetro, por exemplo, um valor máximo da distância euclidiana entre eles para qual os resultados das estimativas sejam aceitáveis e em caso de não satisfação desse critério, utilizar a média dos fatores do grupo. A não inclusão de séries temporais de volume de tráfego restringe a análise à comparação estática entre postos. Além disso, a escolha por um único algoritmo de agrupamento limitou a diversidade de métodos analisados, o que abre caminho para aprofundamentos metodológicos futuros, descritos a seguir.

6.1 Sugestões de trabalhos futuros

- **Explorar algoritmos de agrupamento alternativos:** utilizar métodos como *DBSCAN*, baseado em densidade, ou partições *fuzzy*, que permitem a sobreposição entre grupos. Essas

abordagens podem revelar novas relações entre postos e oferecer comparações importantes ao *K-means*, ampliando a robustez da segmentação.

- **Incluir análise temporal de tráfego:** incorporar séries temporais de volume de tráfego por meio de técnicas como *Dynamic Time Warping* (DTW), que permite comparar curvas de variação no tempo. Essa abordagem pode capturar padrões dinâmicos entre postos e refinar a associação entre eles com base no comportamento ao longo do ano. Principalmente incorporar curvas de variação mensal do volume, por serem dados já presentes para os postos do PNCT.
- **Estimar o volume médio diário anual (VMD_a):** aplicar modelos de regressão, como a regressão linear, para prever o VMD_a em postos de cobertura a partir de variáveis geográficas. Essa modelagem complementar a associação por similaridade e forneceria uma estimativa direta de volume.
- **Aplicar validação estatística dos resultados:** utilizar métricas como *silhouette score* ou análise de variância intra e entre grupos para avaliar a qualidade dos agrupamentos. Isso permitiria uma verificação quantitativa da segmentação e maior confiança nos padrões identificados.
- **Utilizar técnicas de redução de dimensionalidade:** empregar *Principal Component Analysis* (PCA) para identificar combinações latentes de variáveis com maior poder explicativo. A interpretação dos *components* resultantes pode facilitar a compreensão dos fatores que mais diferenciam os grupos formados.

Referências

- Almeida, A.; Carvalho, F.; Menino, F. **Introdução ao Machine Learning**. [S.l.], 2017. Disponível em: <<https://dataat.github.io/introducao-ao-machine-learning/index.html>>.
- Bellamy, P. H. **Seasonal Variations in Traffic Flows**. Berkshire, Great Britain, 1978. Prepared by Department of the Environment and the Department of the Transport.
- Bishop, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer Science+Business Media, 2006.
- CNT. **DNIT libera dados da contagem de tráfego em rodovias federais**. 2016. Acesso em: 26 jun. 2025. Disponível em: <<https://cnt.org.br/agencia-cnt/dnit-libera-dados-da-contagem-de-trafego-em-rodovias-federais-cnt>>.
- CNT; SEST SENAT. **Pesquisa CNT de Rodovias 2023**. Brasília: CNT : SEST SENAT : ITL, 2023. 204 p. Ilustrações coloridas, mapas, gráficos.
- CNT; SEST SENAT. **Pesquisa CNT de Rodovias 2024**. Brasília, 2024. 224 p.: il. color.; mapas, gráficos. ISBN 978-85-68865-24-8 (Impresso).
- DNIT. **Produto 2 – Listagem dos postos de contagem e seus VMD para o ano de 2006**. 2008.
- DNIT. **Estimativa de VMDa 2017 (Metodologia e Resultados)**. 2018. Apresentação técnica. Apresentação realizada por Glaydston Mattos Ribeiro e Saul Germano Rabello Quadros durante o evento 'Segunda Semana do Planejamento' sobre as atividades do DNIT, em junho de 2018. Disponível em: <<https://www.gov.br/dnit/pt-br/assuntos/campanhas-e-eventos/semana-do-planejamento/2a-semana-do-planejamento/4-estimativa-vmda-glaydston-saul.pdf>>.
- DNIT. **Gestão e Disponibilização dos Dados de Tráfego do PNCT**. 2018. Apresentação técnica. Apresentação realizada por Eng. Leonel Antônio da Rocha Teixeira Junior, Analista em Infraestrutura de Transportes, durante evento 'Segunda Semana do Planejamento' sobre as atividades do DNIT.
- DNIT/IPR. **Manual de estudos de tráfego**. Rio de Janeiro, 2006. Departamento Nacional de Infra-Estrutura de Transportes. Diretoria de Planejamento e Pesquisa. Coordenação Geral de Estudos e Pesquisa. Instituto de Pesquisas Rodoviárias. Publicação IPR - 723. 384 p.
- Drusch, R. Estimating annual average daily traffic from short-term traffic counts. **Highway Research Record**, 1966.
- FHWA. **Traffic Monitoring Guide**. Washington, D.C., 2022. Office of Highway Policy Information.
- Han, J.; Kamber, M.; Pei, J. **Data Mining: Concepts and Techniques**. 2. ed. [S.l.]: Elsevier, 2006.
- Jolly, K. **Machine Learning with Scikit-Learn Quick Start Guide: Classification, Regression, and Clustering Techniques in Python**. [S.l.]: Packt Publishing, 2018. Disponível em: <<http://ebookcentral.proquest.com/lib/cqu/detail.action?docID=55978>>.

- McCulloch, W. S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biology**, Pergamon Press plc / Society for Mathematical Biology, v. 52, n. 1/2, p. 99–115, 1943. Reimpressão em 1990 no Bulletin of Mathematical Biology.
- Michilles, R. **Elbow Method e Silhouette Score: como identificar o número ideal de clusters**. 2025. Disponível em: <<https://hub.asimov.academy/blog/elbow-method-e-silhouette-score-numero-cluster/>>.
- Ministério dos Transportes. **Frota de Veículos 2023: Frota Nacional (Dezembro 2023) por município e UF**. 2023. Disponível em: <<https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2023>>. Acesso em: 11 de julho de 2025. Disponível em: <<https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2023>>.
- Müller, A. C.; Guido, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. [S.l.]: O'Reilly Media, 2016.
- Ortúzar, J. de D.; Willumsen, L. G. **Modelling Transport**. 4. ed. [S.l.]: John Wiley & Sons, 2011.
- Papacostas, C. S.; Prevedouros, P. D. **Transportation Engineering and Planning**. 3. ed. Pearson/Prentice Hall, 2009. ISBN 0131973096,9780131973091. Disponível em: <<http://gen.lib.rus.ec/book/index.php?md5=9C4358B2FA33CEC88B3E190364E464EB>>.
- Pignataro, L. J. **Traffic engineering: theory and practice**. [S.l.]: Prentice-Hall, 1973.
- PNCT. **Metodologia de Coleta de Dados – Plano Nacional de Contagem de Tráfego (PNCT)**. 2023. Acesso em: 27 jun. 2025. Disponível em: <<https://servicos.dnit.gov.br/dadospnct/metodologia>>.
- Russell, S.; Norvig, P. **Artificial Intelligence: A Modern Approach**. 4. ed. [S.l.]: Pearson, 2020.
- Schneider, W. H.; Tsapakis, I. **Review of Traffic Monitoring Factor Groupings and the Determination of Seasonal Adjustment Factors for Cars and Trucks**. [S.l.], 2009.
- Shah, S. C.; Hirschmann, J. J. **Analysis of Routine Traffic Stations to Optimize Locations and Frequency**. Louisiana, 1981. Final Report, 31 p.
- Treiber, M.; Kesting, A. **Traffic Flow Dynamics: Data, Models and Simulation**. [S.l.]: Springer, 2013.
- Turing, A. M. Computing machinery and intelligence. **Mind**, v. 49, p. 433–460, 1950.

Anexos

A Códigos de *Python*

A.1 Aquisição de dados

Listing 1: Postos Permanentes

```
1  import requests
2  import geopandas as gpd
3  from shapely.geometry import Point
4  import os
5  import time
6  import numpy as np
7  import sys
8
9  # =====
10 output_dir = r"C:\dados_tcc\tcc_originais\PNCT\gpkg"
11 historico = False # Período de 2014 a 2019: historico=True; Período de 2021 a 2023: historico=False
12 # =====
13
14 # Diretório de saída
15 if not os.path.isdir(output_dir):
16     print(f"Erro: O diretório '{output_dir}' não existe.")
17     sys.exit(1)
18
19 # URL para obter todos os postos
20 url_postos = f"https://servicos.dnit.gov.br/dadospnct-api/api/v1/coletaContinua?historico={historico}"
21 response = requests.get(url_postos)
22 response.raise_for_status()
23 data = response.json()
24
25 geometry = []
26 properties = []
27 id_postos = []
28
29 # Processar os postos e coletar IDs
30 for feature in data["features"]:
31     coords = feature["geometry"]["coordinates"]
32     geometry.append(Point(coords[0], coords[1]))
33
34     prop = feature["properties"]
35     id_postos.append(prop["id"])
36     properties.append(prop)
37
38 print(f'Coletando {len(id_postos)} postos permanentes.\n')
39
40 # Criar GeoDataFrame inicial
41 gdf = gpd.GeoDataFrame(properties, geometry=geometry)
42 gdf.set_crs("EPSG:4326", inplace=True) #Definir o CRS
43
```

```

44 # Adicionar coluna "Tipo de Coleta"
45 gdf["Tipo de Coleta"] = "Permanente"
46
47 # Salvar GeoPackage
48 output_gpkg = f"{output_dir}/pnct_permanente_2021-2023.gpkg"
49 gdf.to_file(output_gpkg, driver="GPKG", layer="pnct") # 'layer=' define o nome da camada dentro do GPKG
50
51 print(f"\nGeoPackage gerado com sucesso: {output_gpkg}")

```

Listing 2: Postos Contínuos

```

1 import requests
2 import geopandas as gpd
3 from shapely.geometry import Point
4 import os
5 import time
6 import numpy as np
7 import sys
8
9 # =====
10 output_dir = r"C:\dados_tcc\tcc_originais\PNCT\gpkg"
11 # Período de 2014 a 2019: historico=True; Período de 2021 a 2023: historico=False
12 historico = False
13 # =====
14
15 # Diretório de saída
16 if not os.path.isdir(output_dir):
17     print(f"Erro: O diretório '{output_dir}' não existe.")
18     sys.exit(1) # Termina o programa com código de erro 1
19
20 # URL para obter todos os postos de cobertura
21 url_postos = f"https://servicos.dnit.gov.br/dadospnct-api/api/v1/coleitaCobertura?historico={historico}"
22 response = requests.get(url_postos)
23 response.raise_for_status()
24 data = response.json()
25
26 geometry = []
27 properties = []
28 id_postos = []
29
30 # Processar os postos e coletar IDs
31 for feature in data["features"]:
32     coords = feature["geometry"]["coordinates"]
33     geometry.append(Point(coords[0], coords[1]))
34
35     prop = feature["properties"]
36     id_postos.append(prop["id"])
37     properties.append(prop)
38
39 print(f'Coletando {len(id_postos)} postos de cobertura.\n')
40
41 # Criar GeoDataFrame inicial
42 gdf = gpd.GeoDataFrame(properties, geometry=geometry)

```

```

43 gdf.set_crs("EPSG:4326", inplace=True) #Definir o CRS
44
45 # Adicionar coluna "Tipo de Coleta"
46 gdf["Tipo de Coleta"] = "Cobertura"
47 print(gdf)
48
49 # Salvar GeoPackage
50 output_gpkg = f"{output_dir}/pnct_cobertura_2021-2023.gpkg"
51 gdf.to_file(output_gpkg, driver="GPKG", layer="pnct") # 'layer=' define o nome da camada dentro do GPKG
52
53 print(f"\nGeoPackage gerado com sucesso: {output_gpkg}")

```

Listing 3: Merge dos Postos

```

1 import geopandas as gpd
2 import pandas as pd
3 import os
4
5 # Diretório onde estão os arquivos
6 output_dir = r"C:\dados_tcc\tcc_originais\PNCT\gpkg"
7
8 # Caminhos dos arquivos de entrada
9 file_permanente = os.path.join(output_dir, "pnct_permanente_2021-2023.gpkg")
10 file_cobertura = os.path.join(output_dir, "pnct_cobertura_2021-2023.gpkg")
11
12 # Caminho do arquivo de saída
13 file_saida = os.path.join(output_dir, "pnct_todos_postos_2021-2023.gpkg")
14
15 # Carregar os GeoDataFrames
16 gdf_permanente = gpd.read_file(file_permanente, layer="pnct")
17 gdf_cobertura = gpd.read_file(file_cobertura, layer="pnct")
18
19 # Concatenar os dois GeoDataFrames
20 gdf_todos = gpd.GeoDataFrame(pd.concat([gdf_permanente, gdf_cobertura], ignore_index=True))
21
22 # Organizar as colunas na ordem desejada
23 colunas_ordenadas = [
24     "id", "uf", "br", "km",
25     "Tipo de Coleta", "geometry"
26 ]
27 gdf_todos = gdf_todos[colunas_ordenadas]
28
29 # Definir CRS (por segurança, se necessário)
30 gdf_todos.set_crs("EPSG:4326", inplace=True)
31
32 # Salvar o resultado no arquivo de saída
33 gdf_todos.to_file(file_saida, driver="GPKG", layer="pnct")
34
35 print(f"Arquivo combinado salvo com sucesso em: {file_saida}")

```

Listing 4: População e Renda per Capita

```

1 import geopandas as gpd
2
3 # =====
4 postos_file_path = r"C:\dados_tcc\tcc_processamento\postos_snv.gpkg"
5 input_municipios = r"C:\dados_tcc\tcc_processamento\ibge\municipios_com_alfa.gpkg" # Malha municipal com
↳ dados censitários
6 raio = 50000 # metros
7 # =====
8
9 # === 1. Carregar as camadas ===
10 postos = gpd.read_file(postos_file_path) # ou .shp
11 municipios = gpd.read_file(input_municipios)
12
13 # === 2. Garantir que estão no mesmo CRS projetado (ex: SIRGAS 2000 / Polyconic) ===
14 postos = postos.to_crs(epsg=5880)
15 municipios = municipios.to_crs(postos.crs)
16
17 # Remove municípios com dados nulos
18 municipios = municipios[
19     municipios["renda_per_capita"].notnull() & municipios["moradores"].notnull()
20 ]
21
22 # === 3. Criar buffers de 50 km (50000 metros) ===
23 postos["geometry_buffer"] = postos.geometry.buffer(raio)
24
25 # === 4. Inicializar colunas de saída ===
26 postos["renda_per_capita"] = 0.0
27 postos["populacao"] = 0.0
28
29 # === 5. Iterar sobre os postos ===
30 for idx, posto in postos.iterrows():
31     buffer_geom = posto["geometry_buffer"]
32
33     # Interseção com municípios
34     intersec = municipios[municipios.intersects(buffer_geom)].copy()
35     intersec["area_total_mun"] = intersec.geometry.area
36     intersec["geometry"] = intersec.geometry.intersection(buffer_geom)
37     intersec["area_intersec"] = intersec.geometry.area
38
39     # Peso da área sobreposta
40     intersec["peso"] = intersec["area_intersec"] / intersec["area_total_mun"]
41
42     # Renda ponderada e população ponderada
43     intersec["renda_ponderada"] = intersec["peso"] * intersec["renda_per_capita"]
44     intersec["populacao_ponderada"] = intersec["peso"] * intersec["moradores"]
45
46     # Soma os valores e salva no posto
47     postos.at[idx, "renda_per_capita"] = round(intersec["renda_ponderada"].sum(), 2)
48     postos.at[idx, "populacao"] = round(intersec["populacao_ponderada"].sum(), 2)
49
50 # === 6. Remover geometria do buffer (opcional)
51 postos.drop(columns=["geometry_buffer"], inplace=True)
52
53 postos = postos.rename(columns={"renda_per_capita": "rpc", "populacao": "pop"})

```

```

54
55 # === 7. Salvar resultado ===
56 postos.to_file(postos_file_path, driver="GPKG")
57 print("Renda e população ponderadas por buffer adicionadas à camada de postos.")

```

A.2 Preparação de dados

Listing 5: *One-Hot Encoding* e Codificação binária de variáveis categóricas

```

1 import pandas as pd
2
3 # Input e Output
4 postos_file_path = r"C:\dados_tcc\tcc_processamento\postos_snv_clean.csv"
5 output_path = r"C:\dados_tcc\tcc_processamento\postos_snv_encoded_2.csv"
6
7 # Carregar o arquivo CSV
8 postos = pd.read_csv(postos_file_path)
9 print(postos)
10 postos = postos.drop_duplicates(subset=[col for col in postos.columns if col != 'fid'])
11 print(postos)
12
13 # Codificação binária para a coluna 'Tipo de Coleta'
14 # A coluna 'Tipo_de_Coleta_Permanente' recebe 1 se for 'Permanente' e 0 se for 'Cobertura'
15 postos['Tipo_de_Coleta_Permanente'] = postos['Tipo de Coleta'].apply(lambda x: 1 if x == 'Permanente'
↵ else 0)
16
17 # Codificação binária para a coluna 'jur' (jurisdição)
18 # A coluna 'jur_Federal' recebe 1 se for 'Federal' e 0 se for 'Estadual'
19 postos['jur_Federal'] = postos['jur'].apply(lambda x: 1 if x == 'Federal' else 0)
20
21 # Codificação da coluna 'sup' usando pd.get_dummies, que cria colunas binárias para cada valor distinto
22 # Converte os valores 'True'/'False' para 1/0
23 postos_encoded = pd.get_dummies(postos, columns=['sup'], drop_first=False)
24
25 # Converte as colunas 'sup' de True/False para 1/0
26 for col in ['sup_DUP', 'sup_IMP', 'sup_PAV', 'sup_PLA']:
27     postos_encoded[col] = postos_encoded[col].astype(int)
28
29 # Salva o resultado em um novo arquivo CSV
30 postos_encoded.to_csv(output_path, index=False)
31
32 # Exibe as primeiras linhas para verificar o resultado
33 print(postos_encoded.head())
34 print(postos_encoded.columns)
35
36 # Verificação de sanidade dos dados
37 # Mostra o maior e menor valor de cada coluna (exceto 'fid' e 'id')
38 for column in postos_encoded.columns:
39     if column not in ['fid', 'id']: # Ignorar 'fid' e 'id'
40         max_value = postos_encoded[column].max()
41         min_value = postos_encoded[column].min()
42         print(f"{column} max = {max_value}")

```

```

43         print(f"{column} min = {min_value}")
44
45     print("Codificação binária aplicada com sucesso e arquivo salvo!")

```

A.3 Agrupamento

Listing 6: Plotagem do método do cotovelo

```

1  import pandas as pd
2  import numpy as np
3  import matplotlib.pyplot as plt
4  from sklearn.cluster import KMeans
5  from sklearn.preprocessing import StandardScaler
6
7  # Caminho para o arquivo CSV
8  file_path = r"C:\dados_tcc\tcc_processamento\postos_snv_encoded_2.csv"
9
10 # Carregar o arquivo CSV com os dados
11 postos = pd.read_csv(file_path)
12
13 # Seleciona as colunas descritivas, a coluna 'ext' e todas as outras à direita
14 # E remove a coluna 'Tipo_de_Coleta_Permanente' que não é usada no agrupamento
15 cols_to_cluster = postos.columns[postos.columns.get_loc('ext'):].drop('Tipo_de_Coleta_Permanente')
16
17 # Filtra os dados apenas com as colunas relevantes para o clustering
18 X = postos[cols_to_cluster]
19
20 # Normaliza os dados para que as variáveis tenham a mesma escala
21 scaler = StandardScaler()
22 X_scaled = scaler.fit_transform(X)
23
24 # Método do Cotovelo (Elbow Method) - Calcula a soma dos erros quadrados dentro do cluster (SSE) para
25 ↪ diferentes números de clusters
26 sse = []
27 range_n_clusters = range(2, 11) # Intervalo de 1 à 10 clusters
28
29 for n_clusters in range_n_clusters:
30     kmeans = KMeans(n_clusters=n_clusters, random_state=42)
31     kmeans.fit(X_scaled)
32     # Inertia (inércia) é a soma dos erros quadrados dentro do cluster (SSE)
33     sse.append(kmeans.inertia_)
34
35 # Plotando o gráfico do método do cotovelo
36 plt.figure(figsize=(8, 6))
37 plt.plot(range_n_clusters, sse, marker='o')
38 plt.title('Método do Cotovelo para Determinar o Número de Clusters')
39 plt.xlabel('Número de Clusters')
40 plt.ylabel('Soma dos Erros Quadrados Dentro do Cluster (SSE)')
41 plt.show()

```

Listing 7: Agrupamento por K-means

```

1  import pandas as pd
2  import numpy as np
3  import matplotlib.pyplot as plt
4  from sklearn.cluster import KMeans
5  from sklearn.metrics import silhouette_score
6  from sklearn.preprocessing import StandardScaler
7  import seaborn as sns
8  import geopandas as gpd
9  import matplotlib.ticker as ticker
10
11  # Caminho para o arquivo CSV e GeoPackage
12  file_path = r"C:\dados_tcc\tcc_processamento\postos_snv_encoded_2.csv"
13  gpkg_file_path = r"C:\dados_tcc\tcc_processamento\postos_snv_clean.gpkg"
14
15  # Carregar o arquivo CSV com os dados
16  postos = pd.read_csv(file_path)
17
18  # Carregar o arquivo GeoPackage com as colunas de geoprocessamento
19  postos_gpkg = gpd.read_file(gpkg_file_path)
20
21  # Selecionar as colunas descritivas, a coluna 'ext' e todas as outras à direita,
22  # e em seguida remover a coluna 'Tipo_de_Coleta_Permanente' que não será usada no agrupamento
23  cols_to_cluster = postos.columns[postos.columns.get_loc('ext'):].drop('Tipo_de_Coleta_Permanente')
24
25  # Filtra os dados apenas com as colunas relevantes para o clustering
26  X = postos[cols_to_cluster]
27
28  # Normalizar os dados para que as variáveis tenham a mesma escala (necessário para K-Means)
29  scaler = StandardScaler()
30  X_scaled = scaler.fit_transform(X)
31
32  # Calcular a curva de Silhouette para determinar o número ideal de clusters
33  silhouette_scores = []
34  range_n_clusters = range(2, 11) # Intervalo de 2 a 10 clusters
35
36  for n_clusters in range_n_clusters:
37      kmeans = KMeans(n_clusters=n_clusters, random_state=42)
38      kmeans.fit(X_scaled)
39      silhouette_avg = silhouette_score(X_scaled, kmeans.labels_)
40      silhouette_scores.append(silhouette_avg)
41      print(f"Silhouette Score para {n_clusters} clusters: {silhouette_avg}")
42
43  # Encontrar o número de clusters com o maior Silhouette Score
44  optimal_n_clusters = range_n_clusters[np.argmax(silhouette_scores)]
45
46  # Plotando a curva de Silhouette
47  plt.figure(figsize=(8, 6))
48  plt.plot(range_n_clusters, silhouette_scores, marker='o')
49  plt.title('Curva de Silhouette para Determinar o Número de Clusters')
50  plt.xlabel('Número de Clusters')
51  plt.ylabel('Silhouette Score')
52  plt.show()
53
54  # Aplica o KMeans com o número ideal de clusters

```

```

55 kmeans = KMeans(n_clusters=optimal_n_clusters, random_state=42)
56 postos['cluster'] = kmeans.fit_predict(X_scaled)
57
58 # Exibe o resultado da clusterização
59 print(postos[['fid', 'cluster']].head())
60
61 # Calcula e exibe os resultados finais do clustering
62 print(f"\nCentroides dos clusters: \n{kmeans.cluster_centers_}")
63 print(f"\nContagem de pontos em cada cluster: \n{postos['cluster'].value_counts()}")
64
65 # === IMPORTÂNCIA DAS VARIÁVEIS NA DEFINIÇÃO DOS CLUSTERS ===
66
67 # Calcula a importância das variáveis com base na proximidade do centroide
68 # As variáveis mais distantes dos centroides são as mais importantes
69 centroids = kmeans.cluster_centers_
70
71 # Normaliza os centroides para compará-los
72 importance = np.abs(centroids).sum(axis=0)
73
74 # Ordena as variáveis pela sua importância (maior para menor)
75 importance_df = pd.DataFrame({'Variable': X.columns, 'Importance': importance})
76 importance_df = importance_df.sort_values(by='Importance', ascending=True)
77
78 # Exibe as variáveis mais importantes
79 print("\nImportância das variáveis na definição do cluster:")
80 print(importance_df)
81
82 # Visualiza a importância das variáveis
83 plt.figure(figsize=(10, 6))
84 plt.barh(importance_df['Variable'], importance_df['Importance'])
85 plt.xlabel('Importance')
86 plt.title('Importância das Variáveis na Definição dos Clusters')
87 plt.show()
88
89 # === VISUALIZAÇÃO DOS CLUSTERS COM CENTROIDES ===
90 plt.figure(figsize=(10, 6))
91
92 # Plot de dispersão dos postos coloridos por cluster
93 sns.scatterplot(
94     data=postos,
95     x='pop',
96     y='acid',
97     hue='cluster',
98     palette='viridis',
99     s=100,
100     alpha=0.6,
101     edgecolor='w',
102     linewidth=0.5
103 )
104
105 # Formataador bonito para o eixo X (população)
106 ax = plt.gca()
107 ax.xaxis.set_major_formatter(ticker.FuncFormatter(lambda x, _: f'{int(x):,}'.replace(",",".")))
108

```

```

109 # Títulos e legendas
110 plt.title('Postos por Cluster')
111 plt.xlabel('População')
112 plt.ylabel('Acidentes')
113 plt.legend(title='Cluster', loc='upper right')
114 plt.tight_layout()
115 plt.show()
116
117 # === ADICIONAR A COLUNA 'CLUSTER' AO ARQUIVO GEO PACKAGE ===
118 # Carregar o arquivo GeoPackage com os postos
119 postos_gpkg['cluster'] = postos['cluster']
120
121 # Salvar o GeoPackage modificado
122 postos_gpkg.to_file(gpkg_file_path, driver='GPKG')
123
124 print("\nArquivo GeoPackage salvo com a coluna 'cluster' adicionada.")

```

A.4 Cálculos sobre os grupos

Listing 8: Cálculo dos valores médios das variáveis por *clusters*

```

1 import pandas as pd
2 import geopandas as gpd
3
4 # Caminho para os arquivos
5 csv_file_path = r"C:\dados_tcc\tcc_processamento\postos_snv_encoded_2.csv"
6 gpkg_file_path = r"C:\dados_tcc\tcc_processamento\postos_snv_clean.gpkg"
7 output_file = r"C:\dados_tcc\tcc_processamento\cluster_averages_clean.csv"
8
9 # Carregar os dados
10 postos = pd.read_csv(csv_file_path)
11 postos_gpkg = gpd.read_file(gpkg_file_path)
12
13 # Corrigir escalas erradas (valores 1000x maiores que o correto)
14 postos['ext'] = postos['ext'] / 1000
15 postos['dist_p'] = postos['dist_p'] / 1000
16 postos['dist_a'] = postos['dist_a'] / 1000
17
18 # Colunas de interesse
19 cols_to_average = ['ext', 'rpc', 'pop', 'relevo', 'pois', 'dist_p', 'dist_a', 'acid',
20                   'frota', 'prod', 'jur_Federal', 'sup_DUP', 'sup_IMP', 'sup_PAV', 'sup_PLA']
21
22 # Relacionar os DataFrames
23 postos_gpkg = postos_gpkg.set_index('id')
24 postos = postos.set_index('id')
25 postos['cluster'] = postos_gpkg['cluster']
26
27 # Calcular médias por cluster
28 cluster_averages = postos[cols_to_average + ['cluster']].groupby('cluster').mean()
29 cluster_averages = cluster_averages.round(3)
30
31 # Salvar com vírgula como separador decimal e ponto e vírgula como separador de colunas

```

```

32 cluster_averages.to_csv(output_file, index=True, sep=';', decimal=',')
33
34 print(f"Médias por Cluster salvas com sucesso em: {output_file}")
35
36
37
38

```

Listing 9: Estabelecimento do Posto Permanente mais Similar a cada Posto de Cobertura

```

1  import pandas as pd
2  import numpy as np
3  from sklearn.preprocessing import StandardScaler
4  from scipy.spatial.distance import cdist
5
6  # Caminho para o arquivo CSV
7  file_path = r"C:\dados_tcc\tcc_processamento\postos_snv_encoded_2.csv"
8
9  # Carregar os dados
10 postos = pd.read_csv(file_path)
11
12 # Selecionar as colunas para medição das distâncias
13 cols_to_measure = postos.columns[postos.columns.get_loc('ext'):].drop('Tipo_de_Coleta_Permanente')
14 X = postos[cols_to_measure]
15
16 # Normalizar os dados
17 scaler = StandardScaler()
18 X_scaled = scaler.fit_transform(X)
19
20 # Adicionar vetores normalizados ao DataFrame original
21 postos['vector'] = list(X_scaled)
22
23 # Filtrar postos válidos
24 postos = postos[postos['id'].notna()]
25
26 # Separar cobertura e permanente
27 postos_cobertura = postos[postos['Tipo de Coleta'] == 'Cobertura'].copy()
28 postos_permanente = postos[postos['Tipo de Coleta'] == 'Permanente'].copy()
29
30 # Resetar índices
31 postos_cobertura.reset_index(drop=True, inplace=True)
32 postos_permanente.reset_index(drop=True, inplace=True)
33
34 # Matrizes de vetores
35 vetores_cobertura = np.vstack(postos_cobertura['vector'].to_numpy())
36 vetores_permanente = np.vstack(postos_permanente['vector'].to_numpy())
37
38 # Distâncias euclidianas
39 dist_matrix = cdist(vetores_cobertura, vetores_permanente)
40
41 # Índices dos pares mais próximos
42 idxs_min = dist_matrix.argmin(axis=1)
43 dists_min = dist_matrix.min(axis=1)

```

```

44
45 # Criar tabela com IDs concatenados (sem o KM)
46 resultados = []
47
48 for i, dist in enumerate(dists_min):
49     row_cov = postos_cobertura.loc[i]
50     row_perm = postos_permanente.loc[idxs_min[i]]
51
52     id_cov = f"{row_cov['id']} (UF: {row_cov['uf']}, BR: {row_cov['br']})"
53     id_perm = f"{row_perm['id']} (UF: {row_perm['uf']}, BR: {row_perm['br']})"
54
55     resultados.append({
56         'id_cobertura': id_cov,
57         'id_permanente': id_perm,
58         'distancia': f"{dist:.6f}".replace('.', ',')
59     })
60
61 # Converter para DataFrame
62 result_df = pd.DataFrame(resultados)
63
64 # Exibir
65 print(result_df)
66 print(f"Total de pares encontrados: {len(result_df)}")
67
68 # Salvar CSV com separador ; e vírgula como separador decimal
69 result_df.to_csv(
70     r"C:\dados_tcc\tcc_processamento\postos_similares.csv",
71     sep=';',
72     index=False,
73     encoding='utf-8'
74 )
75
76 print("Tabela de pares similares salva com sucesso.")
77

```