CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

CONSTRUINDO UMA BASE DE DADOS PARA ESTUDOS DE MOBILIDADE URBANA DA CIDADE DO RIO DE JANEIRO

Vinicius dos Santos Vancellote Almeida

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Rafaelli de Carvalho Coutinho
Diego Moreira de Araujo Carvalho

Rio de Janeiro,
Julho de 2025

Construindo uma base de dados para estudos de mobilidade urbana da Cidade do Rio de Janeiro

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Vinicius dos Santos Vancellote Almeida

Aprovada por:

Presidente, Prof. Diego Moreira de Araujo Carvalho, D.Sc. (coorientador)

Profa. Rafaelli Coutinho, D.Sc. (orientadora)

Prof. Eduardo Soares Ogasawara, D.Sc.

Prof. Rafael Garcia Barbastefano, D.Sc.

Prof. Fabio Andre Machado Porto, D.Sc.

Rio de Janeiro,

Julho de 2025

RESUMO

Construindo uma base de dados para estudos de mobilidade urbana da Cidade do Rio de Janeiro

Vinicius dos Santos Vancellote Almeida

Orientadores:
Rafaelli de Carvalho Coutinho
Diego Moreira de Araujo Carvalho

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

A mobilidade urbana está diretamente relacionada à qualidade de vida nas cidades, especialmente diante do crescimento populacional e dos desafios ambientais. O transporte público desempenha um papel fundamental nesse contexto, mas é frequentemente impactado por eventos climáticos adversos, como chuvas intensas, que podem comprometer sua eficiência. Este trabalho apresenta a construção de uma base de dados integrada e enriquecida para estudos de mobilidade urbana na cidade do Rio de Janeiro. A metodologia desenvolvida realiza o tratamento e enriquecimento dos dados de GPS da frota de ônibus da cidade, integrando informações posicionais, pluviométricas, administrativas e de emissões de gases, a partir de um pipeline automatizado. O resultado do processamento é uma base de dados estruturada em cinco tabelas-resumo (DSTs), cada uma voltada para uma dimensão específica da análise urbana. Como exemplo de usabilidade, foi proposta uma metodologia para identificação de pontos de alagamento durante eventos de chuva, utilizando a técnica de clusterização DBSCAN aplicada sobre observações de ônibus parados em regiões com precipitação moderada a forte. A metodologia apresentada como exemplo foi validada por meio da comparação com registros jornalísticos e mostrou resulta dos promissores para auxiliar o poder público na identificação de gargalos de mobilidade. Todos os dados e *scripts* foram disponibilizados abertamente, promovendo a ciência aberta e o reuso da informação.

Palavras-chave: séries espaço-temporais; mobilidade urbana; banco de dados;

Rio de Janeiro,

Julho de 2025

ABSTRACT

Building a Database for Urban Mobility Studies in the City of Rio de Janeiro

Vinicius dos Santos Vancellote Almeida

Advisors:
Rafaelli de Carvalho Coutinho
Diego Moreira de Araujo Carvalho

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Urban mobility is directly related to the quality of life in cities, especially in the face of population growth and environmental challenges. Public transportation plays a fundamental role in this context but is often impacted by adverse weather events, such as heavy rainfall, which can compromise its efficiency. This work presents the construction of an integrated and enriched database for urban mobility studies in the city of Rio de Janeiro. The proposed methodology processes and enriches GPS data from the municipal bus fleet by incorporating positional, pluviometric, administrative, and vehicular emission information through an automated pipeline developed in Python. The result is a structured database composed of five Data Summary Tables (DSTs), each focusing on a specific dimension of urban analysis. As a use case, a methodology was applied to identify flood-prone areas during rainfall events, using the DBSCAN clustering technique on observations of stationary buses in regions with moderate to heavy precipitation. The methodology was validated through comparison with journalistic records and showed promising results to support public decision-making in identifying mobility bottlenecks. All data and scripts have been made openly available, fostering open science and data reuse.

Key-words: urban mobility; data mining; spatiotemporal series; transport; dataset;database;

Rio de Janeiro,

Julho de 2025

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

A447 Almeida, Vinicius dos Santos Vancellote
Construindo uma base de dados para estudos de mobilidade urbana da cidade do Rio de Janeiro / Vinicius dos Santos Vancellote Almeida. — 2025.
36f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2025.
Bibliografia : f. 34-36
Orientador: Rafaelli de Carvalho Coutinho
Coorientador: Diego Moreira de Araujo Carvalho

1. Banco de dados. 2. Transporte urbano – Rio de Janeiro (RJ). 3. Processamento de dados. 4. Análise de séries temporais. 5. Precipitação (Meteorologia). 6. Sistema de posicionamento global. I. Coutinho, Rafaelli de Carvalho. (Orient.). II. Carvalho, Diego Moreira de Araujo. (Coorient.). III. Título.

CDD 005.3

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

Sumário

I Introdução **1**

II Fundamentação Teórica **4**

II.1 Séries Temporais 4

II.2 Séries Espaço-Temporais 4

II.3 Base de Dados 5

III Trabalhos Relacionados **6**

IV Metodologia **8**

IV.1 Dados originais 8

IV.2 *Pipeline* 10

IV.3 Base de dados Final 16

V Usabilidade **18**

VI Exemplo de Uso **24**

VI.1 Limpeza dos Dados 24

VI.2 Partição Horária 24

VI.3 Clusterização 25

VI.4 Agrupamento dos *clusters* 25

VI.5 Análise do Volume de Chuva 25

VI.6 *Cluster* por logradouros 25

VI.7 Análise comparativa com pontos de alagamento históricos 26

VI.8 Resultados Análise Comparativa 26

VII Conclusão **29**

VII.1 Considerações finais 29

VII.2 Trabalhos Futuros 29

A Zonas Pluviométricas do município do Rio de Janeiro **31**

B Bairros do município do Rio de Janeiro **32**

Referências Bibliográficas 34

Lista de Figuras

II.1 Exemplo de Série Temporal 4

IV.1 Área de abrangência para cada Zona Pluviométrica 9

IV.2 Divisão dos bairros do Município do Rio de Janeiro. 10

IV.3 Etapas do Pipeline e seus descritivos. 11

IV.4 Polígono limite para o processamento dos dados. 12

IV.5 *Dataset* é composto apenas por observações únicas que indiquem uma mudança de estado 12

IV.6 Modelo conceitual da Base de Dados. 16

V.1 Densidade de observações por zona pluviométrica 19

V.2 Média de observações por dia por ano 20

V.3 Distribuição do intervalo entre observações por ano 20

V.4 Média de ônibus por dia por ano 20

V.5 Média de ônibus por dia por mês em 2020 21

V.6 Garagens e Terminais do Município do Rio de Janeiro 21

V.7 Velocidade Média por Turno 22

V.8 Quantidade de Observações por Turno 22

V.9 Quantidade de Observações por Bairros 22

V.10 Emissão de CO por Bairros 23

VI.1 Ônibus presos em ponto de alagamento 27

VI.2 Registro GPS dos ônibus presos no ponto de alagamento da reportagem 27

VI.3 *Cluster* formado pelas observações dos ônibus identificados no ponto de alagamento da reportagem 28

Lista de Tabelas

IV.1 Dados brutos disponibilizados no `Data.Rio` 8

IV.2 Dados brutos disponibilizados no `Alerta Rio` 10

IV.3 Dados brutos posicionais de Bairro e Região Administrativa disponibilizados no `Data Rio` 11

IV.4 Descrição das notações utilizadas 14

IV.5 Average pollutant emission rates from diesel buses[Zhai et al., 2008] 16

IV.6 Descrição dos DSTs 17

VI.1 Categorias de chuva pelo AlertaRio 25

VI.2 Resultado da Análise Comparativa 26

VI.3 Resultado da Análise Comparativa 28

A.1 Zonas Pluviométricas e seus Códigos 31

B.1 Zonas Administrativas e Bairros do Rio de Janeiro 33

Lista de Abreviações

DST	*Data Summary Table*	11, 12, 16, 18, 24, 29
GPS	*Global Positioning System*	1, 2, 5, 6, 7, 8, 18, 29
MOBLAB	*Laboratório de Mobilidade Urbana do CEFET/RJ*	8
OSF	*Open Science Framework*	2
SGBD	*Sistema de Gerenciamento de Banco de Dados*	5
SIURB	*Sistema Municipal de Informações Urbana*	8
SRTM	*Shuttle Radr Topography Mission*	11
VSP	*Vehicle Specific Power*	15

Capítulo I Introdução

O setor de transporte público é de extrema importância para o bem estar e desenvolvimento da sociedade, e o seu monitoramento está diretamente atrelado à qualidade de vida nas grandes cidades. Com isso, um grande volume de dados é gerado diretamente pelos operadores de transporte, aumentando a necessidade de transformar esses dados em informação e inteligência de qualidade para auxiliar nas tomadas de decisão dos agentes públicos que operam esse setor em nossas cidades.

Em 1950 cerca de 30% da população mundial vivia em áreas urbanas, atualmente esse número cresceu para mais da metade e tem a expectativa de chegar a 80% até 2050 [Bettencourt and West, 2010]. As pessoas vêm se movimentando para as cidades atraídas por mais fácil acesso a outras pessoas, trabalhos e oportunidades. Um dos efeitos indiretos dessa movimentação é o congestionamento do trânsito, que influencia a acessibilidade [Christodoulou et al., 2020].

Assim, o transporte público desempenha um papel importante na vida cotidiana da população como uma importante infraestrutura da economia [Huang et al., 2019]. Atualmente, o setor de transporte motorizado é responsável pela emissão de mais de 15% de gases poluentes em todo o mundo, portanto reduzir o tráfego de veículos nas cidades se tornou crucial para limitar os efeitos das urgências climáticas [Verbavatz and Barthelemy, 2019, 2020].

De forma mais generalista, entender a mobilidade nos espaços urbanos é importante para o planejamento do transporte e também para entender os demais processos nas cidades, como congestionamentos e propagação de uma epidemia [Verbavatz and Barthelemy, 2020].

A recente disponibilidade de um volume de dados sem precedentes tornou possível iniciar uma nova "Ciência das Cidades", entretanto, ainda não foi possível definir um modelo simples para explicar os mecanismos dominantes que governam a formação e evolução dos padrões da mobilidade [Louf and Barthelemy, 2014]. Como os congestionamentos variam conforme o tempo e espaço, usar dados com resolução espaço-temporal aumenta significativamente a confiabilidade desses impactos [Christodoulou and Christidis, 2020].

As pesquisas relacionadas ao transporte rodoviário fazem análise de dados coletados, principalmente, por meio do Sistema de Posicionamento Global, do inglês *Global Positioning System* (GPS). O GPS fornece dados de posicionamento geográfico, velocidade e tempo do objeto ao qual está associado.

Quando esse tipo de sensor é usado nas pesquisas de mobilidade urbana, ele está frequentemente

associado a veículos como táxis e ônibus ou a dispositivos móveis de pessoas como celular [Cruz et al., 2018].

Devido ao modelo de dados fornecidos pelo GPS, esses dados são frequentemente tratados e modelados como trajetórias. O modelo de dados de trajetória é composto pela sequência de observações feitas sobre um objeto que se move no espaço ao longo do tempo. Cada observação de uma trajetória inclui informações sobre o tempo e a localização do objeto observado [Cruz et al., 2018]. No entanto, esse modelo de dados implica em uma individualização do objeto móvel ao qual o GPS está associado, ou seja, sua análise se baseia na observação de objetos individuais que se movem no espaço. A agregação espaço-temporal dos dados coletados através de GPS é uma metodologia que consiste em associar as observações a posições geográficas predefinidas. Ao fim da agregação, cada conjunto possui uma série temporal associada com dados específicos da região estudada. Esse modelo de dados favorece uma visão resumida e sistemática dos dados, apresentando a mobilidade sob uma perspectiva diferente. A agregação de séries espaço-temporal resulta em uma redução significativa no volume de dados de trajetórias. Além disso, os dados agregados são menos suscetíveis a ruídos quando comparados aos dados de trajetória, proporcionando maior estabilidade a longo prazo [Cruz et al., 2018].

O objetivo deste trabalho é disponibilizar uma base de dados para estudos de mobilidade no município do Rio de Janeiro. Para isso são utilizados dados dos GPS dos ônibus municipais fornecidos pela Prefeitura. A ideia é contribuir com o planejamento da cidade, provendo inteligência para o poder público tratar de forma mais assertiva a situação da mobilidade da cidade.

Para isso, foi desenvolvido um sistema de processamento de dados em Python, que trata os dados brutos fornecidos pelo GPS dos veículos. Este sistema enriquece os dados com informações administrativas (como bairro, região administrativa, corredor de transporte, etc.), informações sobre o deslocamento (incluindo velocidade, aceleração e altitude), dados sobre ocorrências pluviométricas (como milímetros de chuva), informações sobre as emissões dos veículos (em gramas de monóxido de carbono, dióxido de carbono, etc.), entre outros subprodutos. Os resultados deste processamento foram disponibilizados por meio de um artigo científico de acesso aberto em jornal de grande circulação. Todos os dados processados e seus aprimoramentos estão acessíveis no repositório aberto *Open Science Framework* (OSF). Os conjuntos de dados disponibilizados agregam valor e conhecimento à mobilidade urbana da cidade, pois possibilitam uma série de análises que correlacionam as observações de GPS dos ônibus com o volume de chuva, as localidades e as emissões de gases poluentes.

Além dessa Introdução, o trabalho está organizado em outros cinco capítulos. O Capítulo II apresenta conceitos importantes para o entendimento do problema. O Capítulo III apresenta os trabalhos relacionados ao tema do projeto. No Capítulo IV são apresentados a metodologia aplicada

e a descrição dos dados gerados. O Capítulo V apresenta gráficos e contextos associados a base de dados, o Capítulo VI apresenta um exemplo de usabilidade para a base de dados e o Capítulo VII, as considerações finais.

Capítulo II Fundamentação Teórica

Neste capítulo são apresentados conceitos associados à mineração de dados em modelos de dados adequados ao contexto de transporte e planejamento público. Na Seções II.1 e II.2 são introduzidos aspectos básicos de séries temporais e séries espaço-temporais, respectivamente. Por fim, na Seção II.3 são abordados os conceitos básicos de dados.

II.1 Séries Temporais

Uma série temporal t é uma sequência ordenada de valores no tempo: $t =< t_1, t_2, ..., t_m >, t_i \in \mathbb{R}$, onde cada t_i é um valor, $m = |t|$ é o número de elementos em t e t_m é o valor mais recente em t [Borges et al., 2020].

Uma série temporal pode ser discreta ou contínua. A série é dita discreta quando as observações são feitas em tempos determinados, normalmente equidistantes, e é dita contínua quando os valores são coletados continuamente. Para padronizar análises, as séries temporais contínuas podem ser discretizadas. Um exemplo de série temporal pode ser visto na Figura II.1.

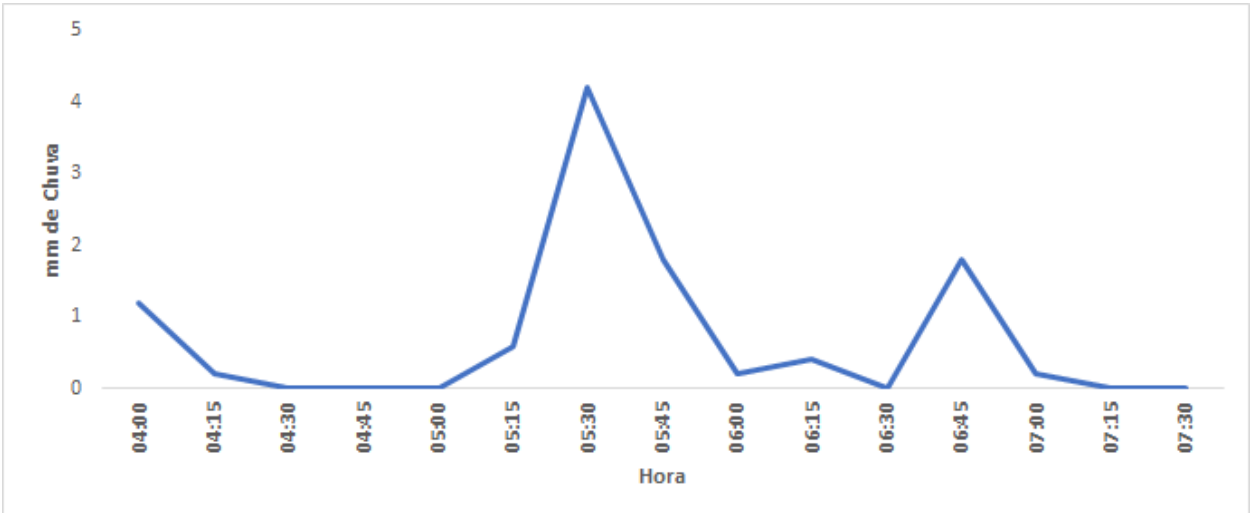


Figura II.1: Exemplo de Série Temporal

II.2 Séries Espaço-Temporais

As séries espaço-temporais são definidas tal como a série temporal com a adição da informação posicional do objeto observado, ou seja, uma série espaço-temporal s é uma série temporal composta

por coordenadas x e y e uma sequência de observações $t =< t_1, t_2, ..., t_m >$. Desta forma, $s.x$ e $s.y$ são coordenadas de s, e $s.t$ é a série de observações de s [Cruz et al., 2018].

As séries espaço-temporais podem ser obtidas por meio de dados emitidos por objetos móveis ou fixos. Os dados de objetos móveis contêm informações de localização que variam com o tempo e normalmente são reportados por meio de GPS, permitindo entender a movimentação desse objeto. Como apontado em Zhang et al. [2020] os dados de GPS são possíveis de obter em uma maior escala, já que podem ser obtidos nos ônibus, táxis ou veículos particulares que circulam durante o dia. Além disso, a partir dos dados de posição em sequência, podemos determinar sua trajetória e estimar sua velocidade, potencializando as análises para essas variáveis.

Os dados de objetos fixos contêm informações em um ponto específico, como por exemplo, um radar de velocidade em que toda vez que um veículo passa, é armazenada sua velocidade, além das informações de data e hora. Tais dados podem ser utilizados para identificar padrões de alterações de velocidade média ao longo do dia, observando as horas críticas com relação a congestionamentos. Cruz et al. [2018] propõe uma tratativa para que dados de trajetória sejam agrupados e tratados como dados de objetos permanentes.

II.3 Base de Dados

Uma base de dados pode ser definida como um conjunto estruturado de dados relacionados, organizado de forma a facilitar seu acesso, gerenciamento e atualização. Segundo Elmasri and Navathe [2011], uma base de dados representa uma coleção de dados inter-relacionados que modelam aspectos relevantes de um domínio do mundo real, sendo projetada para atender a múltiplas aplicações de forma eficiente e sem redundância desnecessária. A estruturação dos dados permite que eles sejam manipulados por sistemas computacionais, viabilizando operações como inserção, consulta, atualização e remoção.

É importante distinguir os conceitos de base de dados e de banco de dados. Date [2004] apresenta que, enquanto a base de dados refere-se ao conjunto de dados propriamente dito, o banco de dados inclui, além da base, o *Sistema de Gerenciamento de Banco de Dados* (SGBD) responsável por controlar o acesso, garantir integridade e prover funcionalidades como concorrência e recuperação de falhas. Assim, a base de dados constitui o núcleo de um sistema de informações, sendo essencial para a tomada de decisões e o suporte a processos organizacionais.

Capítulo III Trabalhos Relacionados

Para a identificação de trabalhos relacionados foi realizada uma busca na base de artigos acadêmicos Scopus, utilizando diferentes *strings* de busca. A *string* "flood"AND "gps"AND ("road"OR "transit") AND ("cluster"), foi identificada como mais adequada, tanto pela quantidade quanto pela relevância dos trabalhos retornados. Essa *string* resultou em 65 publicações. Após a leitura do objetivo e do resumo dos trabalhos encontrados, foram selecionados seis trabalhos que abordavam diretamente os temas "trânsito" e "chuva", alinhados a ideia deste projeto.

Esses trabalhos propõem metodologias distintas para estudar os efeitos das chuvas na mobilidade urbana, como o uso de interpolação espacial, entrevistas com motoristas e análise de trajetórias de táxis ou ônibus. Algumas abordagens fazem uso de técnicas de clusterização para detectar padrões anômalos de deslocamento ou gargalos operacionais.

Kong et al. [2022] Investiga a viabilidade do uso de dados de GPS de táxis para mapear áreas alagadas após eventos de inundação em Shenzhen, China. Os dados são agregados por rodovias e rotas, permitindo a análise da mobilidade urbana antes e depois das enchentes. A metodologia inclui técnicas de interpolação espacial para estimar a extensão das áreas afetadas e validar a precisão do mapeamento em comparação com dados meteorológicos e de sensores ambientais.

Bacero and Fillone [2023] Avalia o impacto dos alagamentos na movimentação de caminhoneiros em Manila, Filipinas. Os dados foram coletados por meio de entrevistas com motoristas e matrizes origem-destino, analisando padrões de deslocamento e os desafios enfrentados devido às condições adversas. A metodologia inclui a identificação de rotas alternativas e a avaliação do tempo adicional necessário para realizar entregas em períodos de chuvas intensas, fornecendo insights para o planejamento logístico em regiões vulneráveis a enchentes.

Jing et al. [2018] explora o papel das bicicletas compartilhadas como complemento ao sistema de transporte em Pequim, China. A análise baseia-se em dados de GPS de táxis, incluindo latitude, longitude e ocupação dos veículos, para identificar padrões de rotas e deslocamento. Técnicas de clusterização são aplicadas para segmentar diferentes fluxos de tráfego e avaliar a influência das bicicletas na redução da dependência do transporte motorizado, contribuindo para estratégias de mobilidade sustentável.

Banik et al. [2020] avalia o tempo de viagem em um corredor de transporte coletivo em Chennai, Índia, com base em dados de GPS de ônibus. A análise considera a variação do tempo de

deslocamento ao longo do dia, identificando pontos críticos de congestionamento. Para isso, são confeccionados gráficos de perfil de tempo de viagem, permitindo a visualização de gargalos no trajeto e a formulação de recomendações para otimizar a eficiência do sistema de transporte público.

Wang and Si [2019] Propõe um modelo para detecção de anomalias no trânsito em Pequim, China, utilizando dados de GPS de 12.712 táxis. O estudo foca na identificação de padrões incomuns de deslocamento que possam indicar congestionamentos inesperados, acidentes ou mudanças na infraestrutura viária. A metodologia inclui a limpeza e pré-processamento dos dados, seguida da aplicação de técnicas de clusterização para detectar variações significativas na mobilidade urbana e sugerir ajustes na gestão do tráfego.

Li et al. [2021] O trabalho investiga os impactos das chuvas intensas no tráfego urbano em Shenzhen, China, utilizando dados combinados de estações meteorológicas e GPS de táxis. O estudo categoriza o volume de chuva e correlaciona seus efeitos com a mobilidade urbana, avaliando variações no tempo de viagem e na distribuição do fluxo de veículos. Técnicas de interpolação espacial são empregadas para mapear as regiões mais afetadas, auxiliando no planejamento de políticas de mitigação para minimizar os impactos das condições climáticas adversas na circulação viária.

Observa-se que há trabalhos que objetivam analisar o impacto da chuva na mobilidade urbana da cidade que se utilizam das mais variadas fontes de dados, no entanto, foi identificado que, apesar do avanço dessas propostas, nenhum dos trabalhos encontrados apresenta a disponibilização estruturada dos dados utilizados, tampouco a publicação de uma base que permita a replicação ou a realização de novos estudos a partir da mesma fonte de dados.

Dessa forma, optamos por seguir com uma abordagem que prioriza a geração e disponibilização aberta de uma base de dados enriquecida com informações posicionais, pluviométricas e estimativas de emissão de gases de efeito estufa. Este diferencial permite que pesquisadores, tomadores de decisão e interessados no tema possam realizar análises futuras com base em dados estruturados, promovendo a ciência aberta e ampliando as possibilidades de estudos sobre mobilidade urbana.

Capítulo IV Metodologia

Nesse capítulo são apresentados os dados brutos originais utilizados para a construção da base de dados, bem como todo o processo de elaboração dessa base, desde a etapa de extração dos dados brutos, a criação de um *pipeline* para tratamento dos dados e a base de dados final. O *pipeline* pode ser definido como uma sequência organizada de etapas e/ou processos aplicados aos dados, desde a sua origem até o formato final desejado, com o objetivo de preparar, transformar, enriquecer e tornar os dados prontos para análise e utilização em modelos.

IV.1 Dados originais

Dados dos ônibus

Os dados originais utilizados nesse trabalho foram obtidos inicialmente por meio de uma parceria iniciada em 2014 pelo *Laboratório de Mobilidade Urbana do CEFET/RJ* (MOBLAB) que coincidiu com a implantação do *Sistema Municipal de Informações Urbana* (SIURB) pela Prefeitura do Rio de Janeiro. O sistema passou a fornecer informações relacionadas ao transporte público da cidade do Rio de Janeiro.

Deste modo, cada veículo do sistema de ônibus da cidade passou a conter uma placa de GPS que emite informações de geolocalização do veículo (LATITUDE e LONGITUDE), bem como seu código de identificação (BUSID), linha de serviço (LINE), velocidade instantânea (VELOCITY) e o registro horário em formato data e hora da observação (GPSTIMESTAMP). Essas informações são coletadas pela Prefeitura e disponibilizadas a partir da plataforma Data.Rio. O descritivo dos dados brutos disponibilizados está apresentado na Tabela IV.1.

Tabela IV.1: Dados brutos disponibilizados no `Data.Rio`

Tipo de dado	Atributo	Descrição
timestamp	GPSTIMESTAMP	Data no formato GMT-0
string	BUSID	Número de ordem do ônibus
string	LINE	Linha de serviço do ônibus
double	LATITUDE	Latitude
double	LONGITUDE	Longitude
double	VELOCITY	Velocidade instantânea fornecida pelo GPS

Inicialmente, não há nenhum tipo de tratamento nos dados, de modo que ao extrair essas in-

formações em um intervalo de tempo sequencial verifica-se uma quantidade relevante de dados duplicados, ou seja, comunicações repetidas que possuem, inclusive, o mesmo registro horário (*timestamp*). Além disso, no início do período de coleta os dados disponibilizados não possuem registro histórico, de modo que caso o *pipeline* de coleta apresentasse algum problema as informações desse período não conseguiriam ser obtidas posteriormente.

Dados pluviométricos

Para beneficiar os dados com informações pluviométricas, foi necessário encontrar registros pluviométricos da cidade do Rio de Janeiro que pudessem ser cruzados com os registros dos ônibus explicados na seção IV.1.

Estes dados foram encontrados na plataforma do Alerta Rio, em que são disponibilizadas as informações de volume de chuva para cada zona pluviométrica do Município em um intervalo de 15 minutos. Nesta plataforma podem-se encontrar registros de chuva desde 1997 e segue atual até os dias de hoje.

Além disso, na plataforma do Data.Rio foi obtida a divisão do Município nas 33 zonas pluviométricas apresentadas na plataforma do Alerta Rio. Essa divisão pode ser vista na Figura IV.1 e a Tabela A.1 apresenta as zonas pluviométricas e seus respectivos códigos. Os descritivos dos dados disponibilizados pluviométricos disponibilizados pelo AlertaRio podem ser vistos na Tabela IV.2.

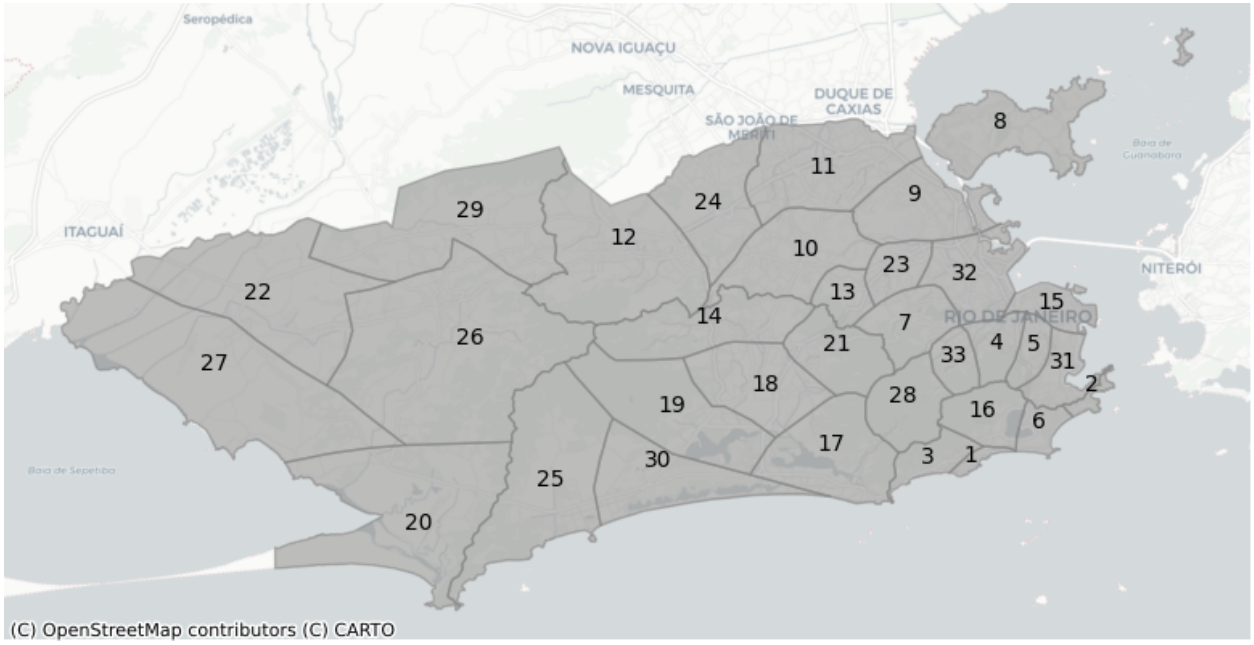


Figura IV.1: Área de abrangência para cada Zona Pluviométrica

Tabela IV.2: Dados brutos disponibilizados no `Alerta Rio`

Tipo de dado	Atributo	Descrição
date	Dia	Data no formato GMT-0
time	Hora	Hora no formato HH:MM:SS
string	HBV	Modelo Hidrológico
double	15 min	Volume acumulado de chuva nos últimos 15 minutos
double	01 h	Volume acumulado de chuva na última hora
double	04 h	Volume acumulado de chuva nas últimas 4 horas
double	24 h	Volume acumulado de chuva nas últimas 24 horas
double	96 h	Volume acumulado de chuva nas últimas 96 horas

Bairros e Regiões Administrativas

De modo similar aos dados pluviométricos, para beneficiar os dados com informações a respeito da localidade em que aquele registro ocorreu, foi utilizada a base disponibilizada no Data.Rio em que é apresentada a divisão do Município do Rio de Janeiro e seus bairros, em formato GeoJson. Essa divisão pode ser vista na Figura IV.2.

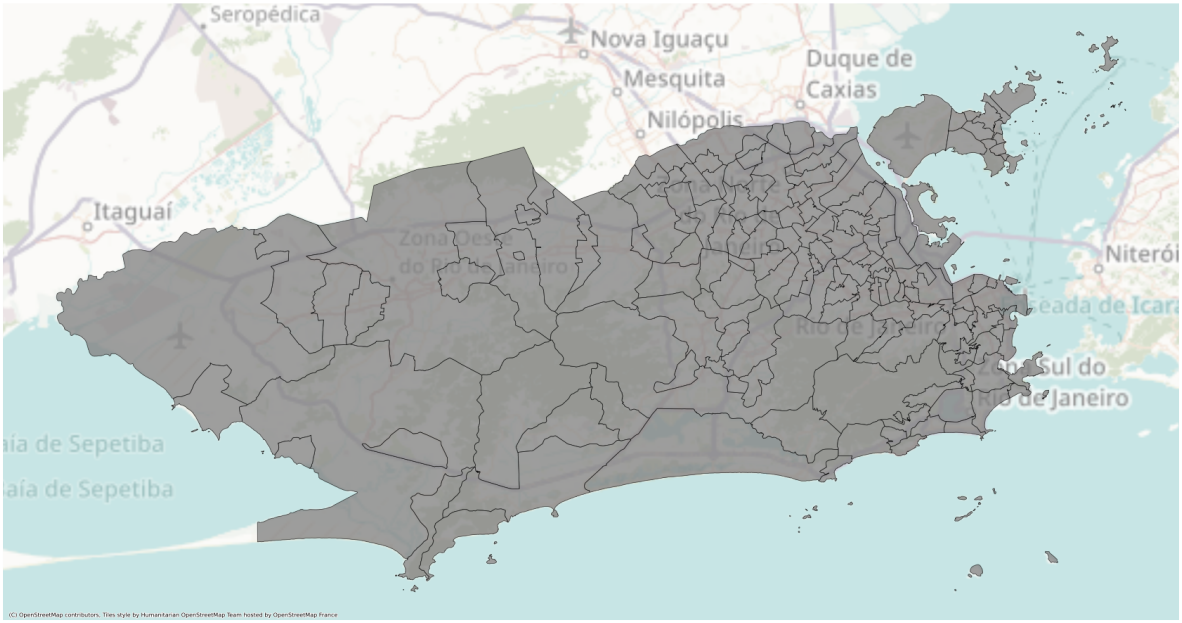


Figura IV.2: Divisão dos bairros do Município do Rio de Janeiro.

Os dados brutos disponibilizados estão descritos na Tabela IV.3.

IV.2 *Pipeline*

Para se obter a base de dados, que é o principal objetivo deste estudo, foi elaborado um *pipeline* de tratamento e geração do dados. Esse *pipeline* foi dividido em 5 etapas, onde ao final de cada uma

Tabela IV.3: Dados brutos posicionais de Bairro e Região Administrativa disponibilizados no `Data Rio`

Tipo de dado	Atributo	Descrição
integer	OBJECTID	Código único do Bairro
float	Área	Tamanho da área do bairro em m²
string	NOME	Nome do Bairro
string	REGIAO_ADM	Nome da Região Administrativa
string	AREA_PLANE	Nome da Região Administrativa
string	CODBAIRRO	Código único do Bairro em formato de 3 dígitos (com zeros a esquerda)
string	CODRA	Código único da Região Administrativa
integer	CODBNUM	Código único do Bairro
string	LINK	Código concatenado utilizando o Nome o código do Bairro
string	RP	Região de Planejamento do Bairro
string	Cod_RP	Código da Região de Planejamento do Bairro
integer	CODBAIRRO_LONG	Código único do Bairro
float	SHAPESTArea	Tamanho da área do bairro em m²
float	SHAPESTLength	Comprimento do perímetro do bairro em m
geometry	geometry	Atributo geométrico do bairro

dessas etapas é gerado um *Data Summary Table* (DST), em que constam as informações obtidas nessa etapa. As cinco etapas estão resumidas na Figura IV.3.

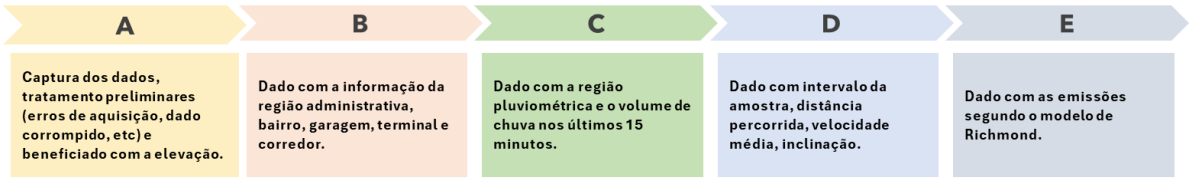


Figura IV.3: Etapas do Pipeline e seus descritivos.

Etapa A

Nessa etapa são extraídos e processados, a cada minuto, os dados originais dos ônibus do *dataset* disponibilizado pela Prefeitura, citados na seção anterior, e armazenados em um servidor do LINCE, de modo que são carregados 1440 arquivos brutos por dia.

Após o processamento dos dados, são feitos alguns tratamentos de limpeza dos dados em que são retirados das bases as observações duplicadas e observações que possuam velocidade instantânea negativa ou acima de 120km/h. Nessa etapa também, é acrescida a informação de elevação, proveniente do *Shuttle Radr Topography Mission* (SRTM)[1], às observações. Para isso, são retirados dados que estão além da área de mapeamento de elevações fornecidos pelo estudo e é traçado um polígono limite, que pode ser visualizado na Figura IV.4, definido pelas seguintes coordenadas: la-

[1] `https://www.earthdata.nasa.gov/data/instruments/srtm`

titude máxima e mínima iguais a -22,6º e -23,2º, respectivamente, e longitude máxima igual -43,0º e mínima igual a -44,0º. Desta forma, é garantido que todas as observações analisadas terão a informação de elevação.



Figura IV.4: Polígono limite para o processamento dos dados.

No fim desta etapa, apenas observações únicas, ou seja, que indiquem uma mudança de estado, estão presentes no DST, conforme explanado na Figura IV.5. Essa etapa do *pipeline* é a única onde são feitas limpezas e tratamentos nos dados, sendo essencial para a confiabilidade do estudo.

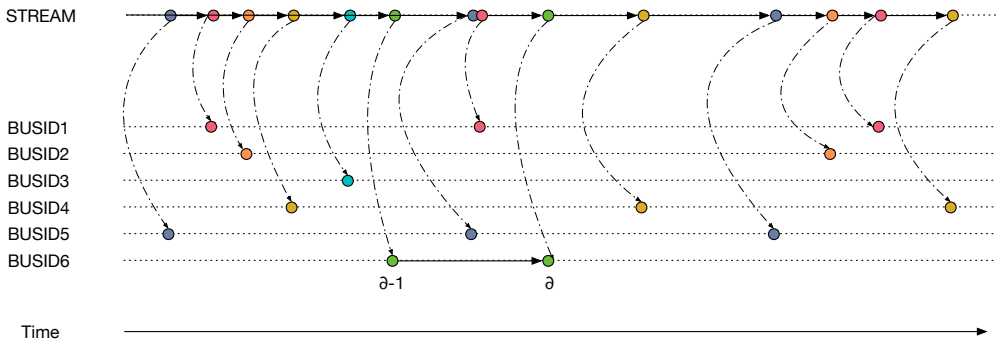


Figura IV.5: *Dataset* é composto apenas por observações únicas que indiquem uma mudança de estado

Etapa B

Nessa etapa são acrescidas as informações de bairro, região administrativa, garagem, terminal e corredor expresso. Para incluir as informações do bairro e região administrativa que a observação está inserida, é utilizado um arquivo GeoJSON disponibilizado pela Prefeitura que divide o Município nos seus bairros.

Para obter as informações das localidades das garagens e terminais de ônibus, foram executadas

técnicas de clusterização nas observações que aconteciam durante a madrugada em que os ônibus possuíam velocidades iguais a zero. Posteriormente foi executada uma análise comparativa com as informações disponibilizadas pela Prefeitura, além de uma análise visual com imagens de satélites. Desta forma, foi gerado um arquivo GeoJSON para que o *pipeline* pudesse identificar se a observação está contida em alguma garagem ou terminal.

As informações das localidades de corredor expresso também são disponibilizadas pela Prefeitura e de forma similar às demais informações, é gerado um GeoJSON que serve como entrada para o *pipeline*.

Essa etapa possibilita que as observações que ocorrem fora da área município sejam identificadas pelo valor nulo nos atributos de bairro e região administrativa. De forma similar, as observações que possuem valor nulo para os atributos de garagem e terminal representam as observações que ocorreram em veículos que estão operacionais e que são produtos deste estudo.

Etapa C

Nessa etapa, de forma muito similar ao que ocorre na etapa anterior, o *pipeline* acrescenta aos dados a informação da zona pluviométrica em que a observação está contida. Para isso, é utilizado um arquivo GeoJSON disponibilizado pela Prefeitura que contém a divisão do município nas suas 33 zonas pluviométricas, conforme apresentado na Figura IV.1. Além disso, o *pipeline* retorna o valor do volume de chuva naquela zona pluviométrica nos últimos 15 minutos, conforme fornecido pelo *dataset* do Alerta Rio.

Etapa D

Nessa seção são analisadas as observações por ônibus, individualmente, como mostrado na Figura IV.5. Para cada sequência de observação δ, são acrescidos os atributos com base na observação anterior, $(\delta - 1)$. Cada valor encontrado é atribuído a observação δ.

Os atributos acrescidos incluem a distância em duas dimensões, hd_δ (a distância Haversine entre duas observações em sequência $\delta - 1$ e δ), a diferença de elevação, h_δ, que é definida usando a diferença de elevações e_δ e $e_{\delta-1}$, o intervalo, i_δ, entre duas observações, a distância tridimensional entre duas observações, d_δ, e a velocidade média, aqui definida como *speed*, s_δ. A Tabela IV.4 descreve as notações utilizadas.

A Equação IV.1 define a distância bi-dimensional, hd_δ, percorrida entre a sequência de observações δ e $\delta - 1$, calculada pela fórmula Haversine (*i.e.* distância entre dois pontos ao longo de uma esfera), em que r é o raio da Terra, $\theta_{\delta-1}$ e θ_δ são as latitudes e $\lambda_{\delta-1}$ e λ_δ são as longitudes das observações

Tabela IV.4: Descrição das notações utilizadas

Notação	Descrição	Atributos
δ	Sequência de observações	ID
t_δ	Registro horário da observação δ	GPSTIMESTAMP
e_δ	Elevação via sensoriamento espacial da observação δ	ELEVATION
dh_δ	Distância bidimensional entre as observações δ e $\delta - 1$	HAVERSINEDISTANCE
h_δ	Diferença de elevação entre as observações e_δ e $e_{\delta-1}$	HEIGHT
i_δ	Intervalo horário entre as observações δ e $\delta - 1$	INTERVAL
d_δ	Distância tridimensional entre as observações δ e $\delta - 1$	DISTANCE
s_δ	Velocidade média entre as observações δ e $\delta - 1$	SPEED
a_δ	Aceleração ou desaceleração do ônibus associado a observação δ	n/a
VSP_δ	Potência Específica do Veículo (*Vehicle specific power*) usada para estimar a emissão média dos gases poluentes CO, CO_2, NO_x, e HC do ônibus associado a observação δ	CO, CO2, NOx, HC

$$hd_\delta = 2r \arcsin\left(\sqrt{sin^2\left(\frac{\theta_\delta - \theta_{\delta-1}}{2}\right) + \cos\left(\theta_{\delta-1}\right)\cos\left(\theta_\delta\right)\sin^2\left(\frac{\lambda_\delta - \lambda_{\delta-1}}{2}\right)}\right) \tag{IV.1}$$

A diferença entre elevações de duas observações subsequentes, h_δ, está descrita na Equação IV.2.

$$h_\delta = e_\delta - e_{\delta-1} \tag{IV.2}$$

O intervalo, i_δ, refere-se a diferença temporal entre a observação atual de um ônibus em particular e a sua observação anterior. Para corrigir longos intervalos sem observação, foi definido um limite de 240 segundos, a partir de uma análise do intervalo médio e seus limites máximos. Se o intervalo for menor ou igual a 240 segundos, a equação que o define é a Equação IV.3, caso contrário, é definido pela Equação IV.4 visto que um intervalo acima de 240 segundos é entendido como uma falha na comunicação do sistema. O resultado, então é atribuído como `NA` (*not available*).

$$i_\delta = t_\delta - t_{\delta-1}, \qquad i_\delta \leq 240s \tag{IV.3}$$

$$i_\delta = \texttt{NA}, \qquad i_\delta > 240s \tag{IV.4}$$

A distância tridimensional percorrida entre duas observações sequenciais, d_δ, é calculada utilizando o Teorema de Pitágoras, definido pela Equação IV.5.

$$d_\delta = \sqrt{hd_\delta^2 + h_\delta^2} \tag{IV.5}$$

A velocidade média, s_δ, é obtida dividindo a distância tridimensional (d_δ) pelo intervalo (i_δ), conforme apresentado na Equação IV.6.

$$s_\delta = \frac{d_\delta}{i_\delta} \tag{IV.6}$$

Etapa E

Na última etapa do *pipeline* são inseridas a base de dados as informações referentes à emissão dos gases poluentes CO, CO_2, HC e NO_x para cada observação. Foi adotada uma abordagem *bottom-up* para calcular as emissões, que são essenciais para o desenvolvimento de políticas públicas.

No entanto, destaca-se que a metodologia empregada não realiza uma análise abrangente das fontes de emissão considerando a natureza e o tipo de atividade responsável pelas emissões. Para análise, utilizamos apenas os ônibus que circulam no Município como fonte de emissões, aplicando o modelo de Potência Específica do Veículo, *Vehicle Specific Power* (VSP), para estimar as taxas de emissão e associá-las às taxas médias de emissão dos ônibus de transporte movidos a diesel, como são os que circulam no Rio de Janeiro [Zhai et al., 2008].

O modelo VSP define a potência por unidade de massa da fonte (kW/ton), sendo uma maneira conveniente de estimar as emissões de um veículo. O modelo considera vários fatores, como aceleração do veículo, inclinação da via, resistência ao rolamento dos pneus e resistência aerodinâmica, como fatores que influenciam a potência absoluta [Koupal et al., 2002]. Uma limitação dessa estimativa é que o peso do veículo afeta as emissões de veículos pesados a diesel Clark et al. [2002], e a contagem de passageiros na frota não está disponível.

O VSP é calculado com base na velocidade do ônibus s_δ, na aceleração ou desaceleração do veículo a, e na inclinação da via $\sin(\alpha) = h_\delta/d_\delta$. Primeiramente, calculamos a aceleração a usando a Equação IV.7, em que $s_\delta - s_{\delta-1}$ representa a variação de velocidade, ou seja, a diferença entre a velocidade observada no momento atual e a observada anteriormente, e o intervalo i_δ, que corresponde à diferença entre o tempo de observação atual e o anterior:

$$a_\delta = \frac{s_\delta - s_{\delta-1}}{i_\delta}. \tag{IV.7}$$

Para duas observações sucessivas, o VSP_δ é calculado usando a Equação IV.8. Em seguida, as taxas médias modais de emissão de CO_2, CO, NO_x e HC são obtidas a partir da Tabela IV.5.

$$VSP_\delta = s_\delta \left[(9.81 a_\delta \frac{h_\delta}{d_\delta} + 0.092) + 0.00021 s_\delta^2 \right] \tag{IV.8}$$

A base de dados fornece as estimativas calculadas em massa por segundo, desta forma é ne-

Tabela IV.5: Average pollutant emission rates from diesel buses[Zhai et al., 2008]

		Taxa média de emissão			
Faixa (bins) VSP	classe VSP	CO_2 (g/s)	CO (g/s)	NO_x (g/s)	HC (mg/s)
$VSP \leq 0$	1	2.4	0.009	0.04	1.23
$0 < VSP < 2$	2	7.8	0.036	0.13	1.7
$2 \leq VSP < 4$	3	12.5	0.045	0.18	1.75
$4 \leq VSP < 6$	4	17.1	0.072	0.22	1.84
$6 \leq VSP < 8$	5	21.2	0.085	0.24	1.94
$8 \leq VSP < 10$	6	24.8	0.091	0.26	2.05
$10 \leq VSP < 13$	7	27.6	0.084	0.28	2.08
$VSP \geq 13$	8	29.5	0.062	0.31	2.15

cessário multiplicar essas taxas de emissão pelo intervalo entre as duas observações para obter a massa total de poluente emitidos naquela movimentação entre as observações.

IV.3 Base de dados Final

Ao final de toda a execução do *pipeline* são gerados 5 DSTs individuais que possuem, além dos atributos já detalhados na seção anterior, um ID exclusivo para cada observação e um atributo com a versão do *pipeline* que gerou os arquivos. A partir do ID exclusivo é possível combinar os DSTs desejados para obter as informações necessárias para um objeto de estudo específico. A descrição de cada DST e seus conteúdos pode ser observada na Tabela IV.6 e o seu modelo conceitual pode ser observado na Figura IV.6.

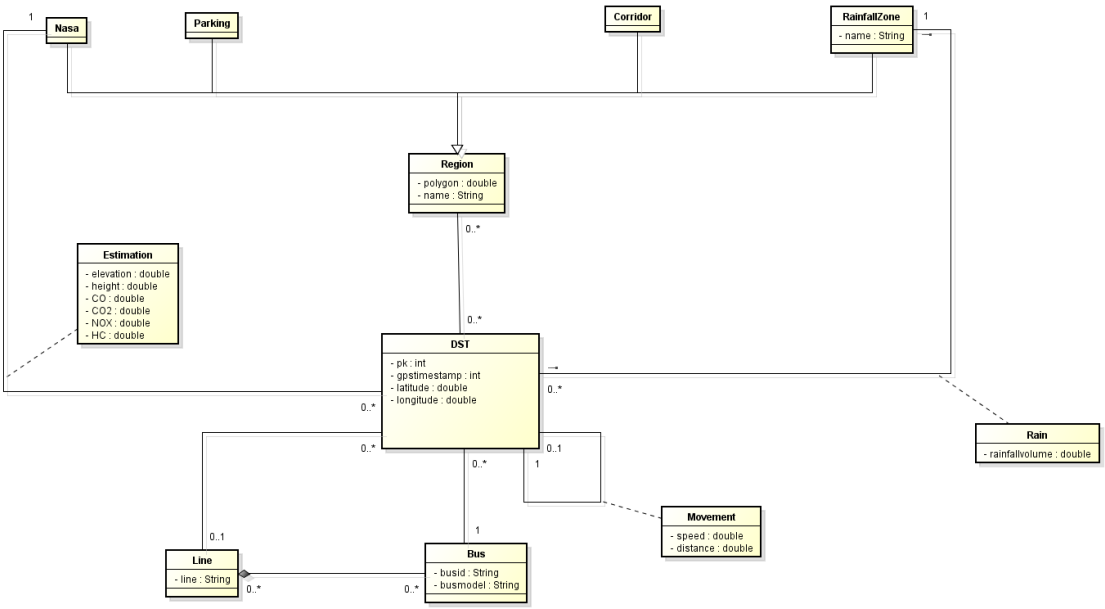


Figura IV.6: Modelo conceitual da Base de Dados.

A produção da base de dados permite uma série de estudos sobre mobilidade urbana, bem como permite auditar e controlar as operadoras de ônibus do Município com relação a oferta e trajeto percorrido pelos veículos.

Tabela IV.6: Descrição dos DSTs

DST-A

Tipo de Dado	Atributo	Descrição
index	ID	Chave Primária
timestamp	GPSTIMESTAMP	Data no formato GMT-0
string	BUSID	Número de ordem do ônibus
string	LINE	Linha de serviço do ônibus
double	LATITUDE	Latitude
double	LONGITUDE	Longitude
double	VELOCITY	Velocidade instantânea emitida pelo GPS (m/s)
double	ELEVATION	Elevação da posição do ônibus extraída a partir dos dados do SRTM (*Shuttle Radar Topography Mission*) [Farr et al., 2007]
string	SWVERSION	Versão do pipeline

DST-B

Tipo de Dado	Atributo	Descrição
index	ID	Chave estrangeira (mesmo ID do DST-A)
string	ADMINISTRATIVERE-GION	ID da Região Administrativa
string	NEIGHBORHOOD	ID do Bairro
double	PARKING	ID da garagem de ônibus
double	TERMINAL	ID do terminal de ônibus
double	CORRIDOR	ID do corredor expresso de ônibus
string	SWVERSION	Versão do pipeline

DST-C

Tipo de Dado	Atributo	Descrição
index	ID	Chave estrangeira (mesmo ID do DST-A)
string	RAINFALLZONE	ID da Zona Pluviométrica
string	RAINFALLVOLUME	Volume de chuva acumulado nos últimos 15 minutos
string	SWVERSION	Versão do pipeline

DST-D

Tipo de Dado	Atributo	Descrição
index	ID	Chave estrangeira (mesmo ID do DST-A)
double	HAVERSINEDISTANCE	Distância bidimensional entre duas observações (m)
double	HEIGHT	Diferença da elevação SRTM entre duas observações (m)
timestamp	INTERVAL	Intervalo entre duas observações (s)
double	DISTANCE	Distância tridimensional entre duas observações (m)
double	SPEED	Velocidade média entre duas observações (m/s)
string	SWVERSION	Versão do pipeline

DST-E

Tipo de Dado	Atributo	Descrição
index	ID	Chave estrangeira (mesmo ID do DST-A)
double	CO	Estimativa da massa de monóxido de carbono emitido entre duas observações (g/s)
double	CO2	Estimativa da massa de dióxido de carbono emitido entre duas observações (g/s)
double	NOx	Estimativa da massa de monóxido de nitrogênio emitido entre duas observações (g/s)
double	HC	Estimativa da massa de hidrocarboneto emitido entre duas observações (g/s)
string	SWVERSION	Versão do pipeline

Capítulo V Usabilidade

A base de dados é composta por 3.228 arquivos estruturados como tabelas-resumo de dados DST no formato Parquet, contendo mais de 9 bilhões de observações distribuídas em 25 atributos, detalhados na Tabela IV.6. Esses arquivos são categorizados em cinco tipos de arquivos DST, que representam diferentes temáticas: informações de posicionamento (DST-A), dados administrativos da cidade (DST-B), dados pluviométricos (DST-C), informações de deslocamento (DST-D) e emissões (DST-E). Um atributo, compartilhado entre todos os DSTs, identifica exclusivamente as observações e correlaciona os DSTs. No DST-A, o atributo ID atua como chave primária; já nos DST-B, DST-C, DST-D e DST-E, ele funciona como chave primária e, ao mesmo tempo, como chave estrangeira referenciando o DST-A.

No DST-A, dois atributos identificam o veículo e a linha de serviço segundo o cadastro da Prefeitura do Rio de Janeiro, permitindo identificar os ônibus observados e monitorar a consistência de suas rotas com as linhas operacionais.

Os atributos são calculados com base em combinações de tempo, espaço e métricas relacionadas ao ambiente e à mobilidade. Os atributos temporais incluem o *timestamp* do GPS do ônibus, permitindo rastrear observações ao longo do tempo. Os atributos espaciais englobam latitude, longitude, elevação, região administrativa, bairro, estacionamento, terminal, corredor e zona de chuva. Um dos atributos representa o volume de chuva acumulado nos últimos 15 minutos. Seis atributos estão associados à mobilidade do ônibus, como distâncias em 2D e 3D, elevação, intervalo, velocidade instantânea e velocidade média. Por fim, quatro atributos estimam as emissões dos gases CO, CO_2, HC e NO_x. Vale destacar que todos os ônibus do Rio de Janeiro utilizam um combustível à base de diesel chamado ARLA 32 (Agente Redutor Líquido Automotivo).

A base de dados possui um arquivo DST para cada dia entre 16 de abril de 2014 e 30 de junho de 2023, totalizando 3.362 arquivos. Nesse período, 134 dias não possuem informações devido a problemas na aquisição dos dados, o que representa menos de 4% do total. As causas da ausência de dados incluem: (i) instabilidade no servidor da Prefeitura que disponibiliza os dados; (ii) falha no servidor de coleta; (iii) interrupções temporárias por conta de manutenções; e (iv) corrupção dos dados durante a geração no servidor.

A Prefeitura do Rio de Janeiro divide a cidade em Zonas de Chuva, onde são instalados e monitorados sensores pluviométricos, conforme explanado na Seção IV.1. A Figura V.1 apresenta

essas zonas, demonstrando a densidade de observações por zona pluviométrica. As áreas próximas à Baía de Guanabara, que compõem a Zona Central da cidade, apresentam maior número de observações do que áreas da Zona Oeste, refletindo o padrão de deslocamento diário da população que se desloca das zonas residenciais para o centro para trabalhar.

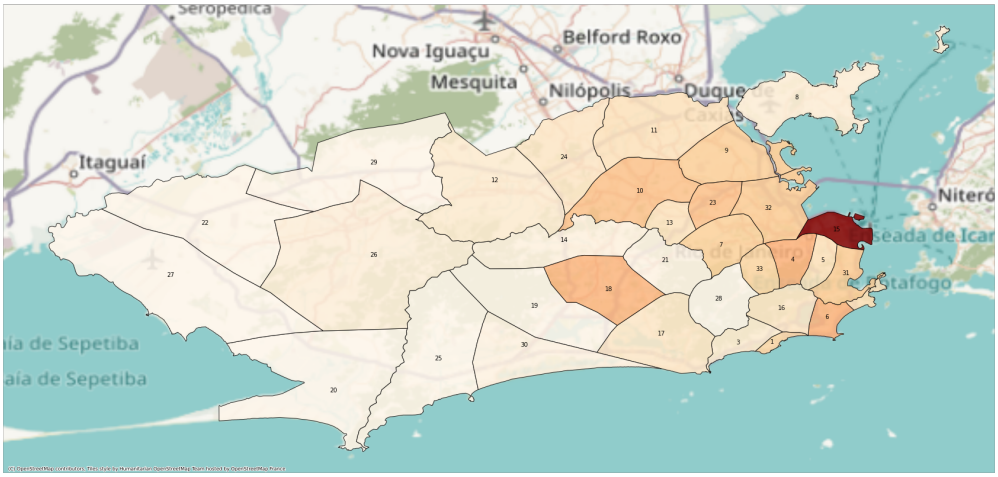


Figura V.1: Densidade de observações por zona pluviométrica

A Figura V.2 mostra a média de observações diárias por ano. O pico foi registrado em 2015, com uma média superior a 4 milhões de observações por dia. Esse número caiu drasticamente para cerca de 2 milhões em 2020, provavelmente devido à pandemia de COVID-19. A Figura V.3 apresenta o intervalo médio entre observações por ano, com aumento de dispersão em 2020 (pandemia) e redução em 2023 após melhorias nos sistemas da Prefeitura. A Figura V.4 exibe a distribuição do número médio de ônibus em circulação por dia ao longo dos anos, também apresentando queda — especialmente em 2020, coincidindo com o início da pandemia, como mostra o limite inferior do *boxplot*. A Figura V.5 reforça essa tendência, evidenciando o declínio de ônibus em circulação a partir de março de 2020, com o endurecimento das medidas de lockdown no Brasil.

Para disponibilizar as informações de garagens e terminais, também foram mapeados os locais de terminais e garagens no Município do Rio de Janeiro. Para isso, foram aplicadas técnicas de clusterização nas observações com velocidade zero, de modo a identificar regiões rotineiras de aglomeração que poderiam indicar a presença de pontos de garagem ou terminais. Após esse mapeamento, foi realizado o cruzamento dessas áreas com informações disponibilizadas pela Prefeitura. As áreas de terminais e garagens podem ser vistas na Figura V.6.

Para analisar as velocidades médias, são considerados apenas os ônibus em circulação, excluindo aqueles estacionados ou nos terminais. Destaca-se que os turnos da noite e da madrugada apresentam as maiores velocidades médias, o que está de acordo com a expectativa, já que o tráfego urbano é menor nesses períodos. A Figura V.7 mostra a velocidade média dos ônibus por turno (madrugada, manhã, tarde e noite) em cada ano, conforme a segmentação da Tabela 8. A Figura

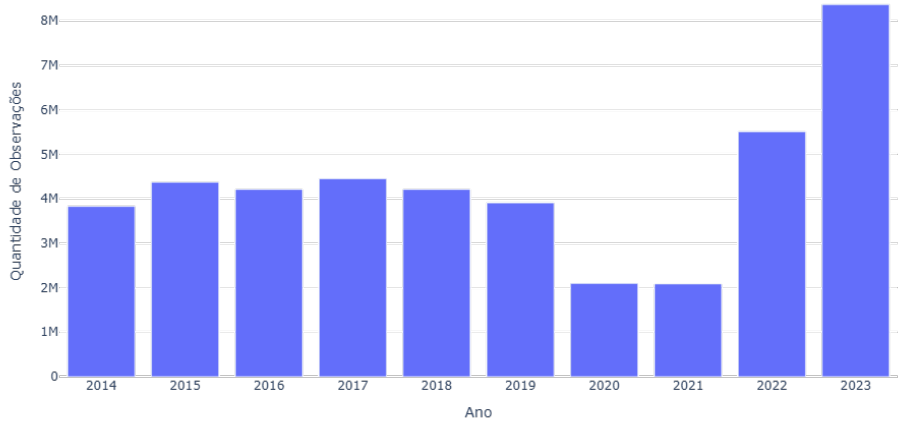


Figura V.2: Média de observações por dia por ano

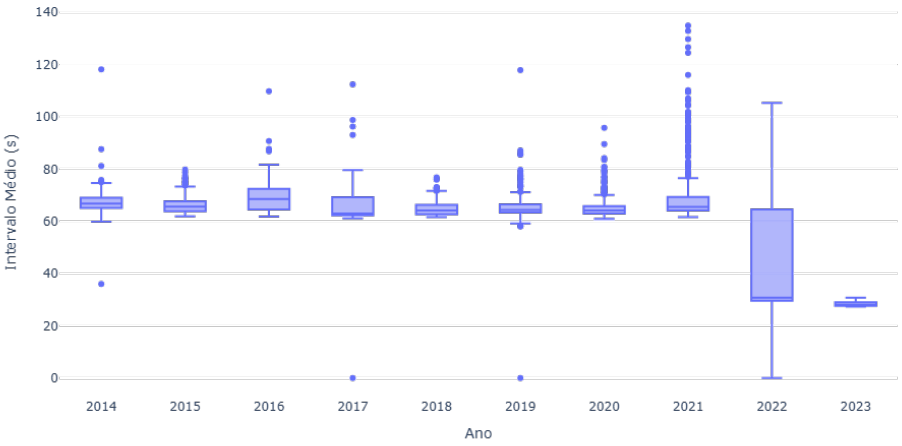


Figura V.3: Distribuição do intervalo entre observações por ano

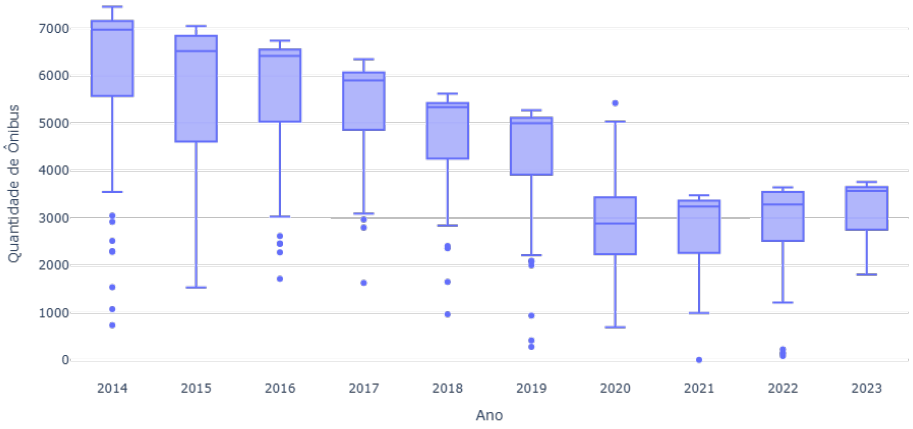


Figura V.4: Média de ônibus por dia por ano

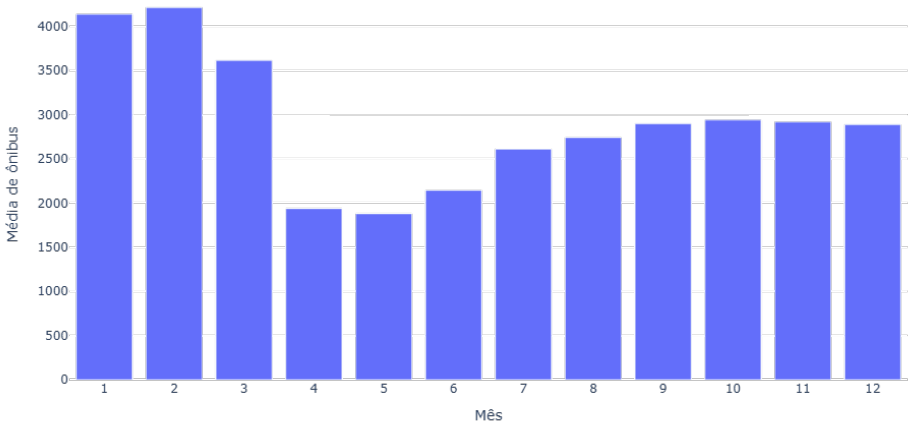


Figura V.5: Média de ônibus por dia por mês em 2020



Figura V.6: Garagens e Terminais do Município do Rio de Janeiro

V.2 mostra o número médio de observações por turno ao longo dos anos.

Analisando os atributos posicionais dos bairros da cidade, temos a Figura V.9 que apresenta um mapa de calor com o total de observações por bairro. Alguns bairros como Freguesia/Jacarepaguá (ID:120), Campo Grande (ID:144) e Centro (ID:5) se destacam com maior número de ônibus circulando. Já a Figura V.10 apresenta o mapa de calor da massa de CO emitida em cada bairro ao longo do período de avaliação, evidenciando regiões com potencialmente pior qualidade do ar, mesmo em áreas com boa ventilação natural.

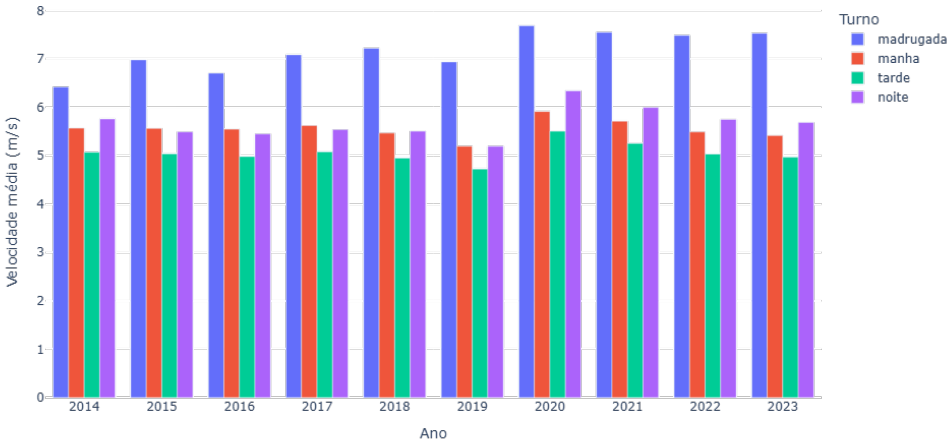


Figura V.7: Velocidade Média por Turno

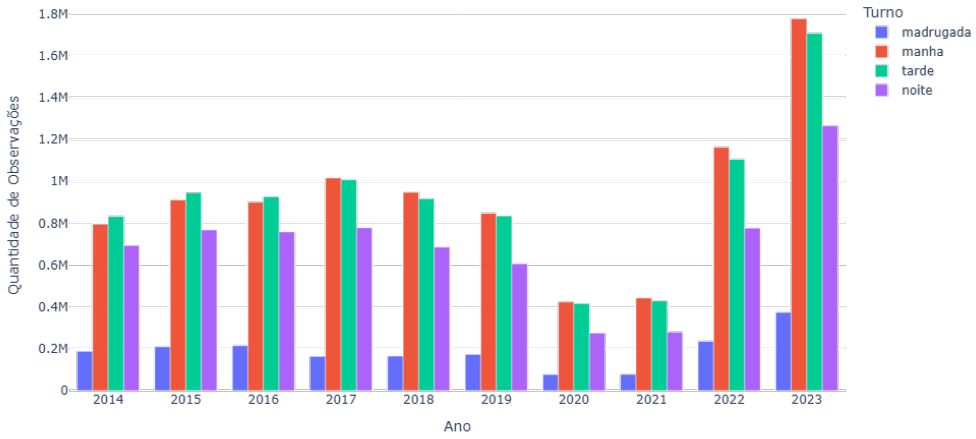


Figura V.8: Quantidade de Observações por Turno

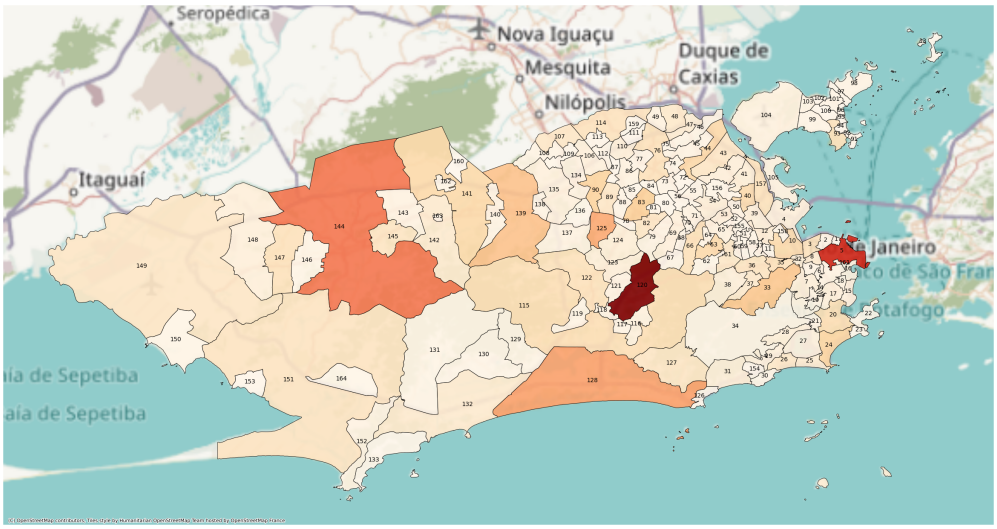


Figura V.9: Quantidade de Observações por Bairros

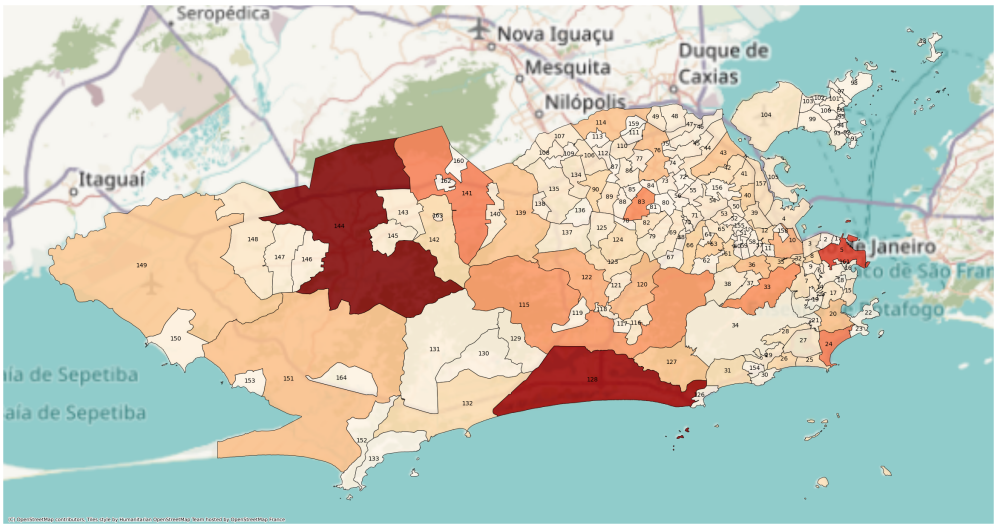


Figura V.10: Emissão de CO por Bairros

Capítulo VI Exemplo de Uso

Devido à complexidade dos dados, é possível executar uma série de análise que agregam conhecimento ao poder público na tomada de decisão. Nesse capítulo é apresentado um exemplo de usabilidade dos dados, cujo objetivo é utilizar os dados para identificar pontos de alagamento em momentos de chuva no Município do Rio de Janeiro, utilizando técnicas de clusterização com DBSCAN. A metodologia se divide nas seguintes etapas: limpeza dos dados, partição horária, clusterização, agrupamento dos *clusters*, análise do volume de chuva, *cluster* por logradouro e análise comparativa com pontos de alagamento históricos. No fim desse capítulo, é apresentado o resultado da metodologia proposta, com o objetivo de confirmar a usabilidade dos dados.

VI.1 Limpeza dos Dados

Para aplicar o DBSCAN foram escolhidos os dados dos DSTs gerados pelo *pipeline*. Inicialmente são necessários apenas os dados posicionais (latitude e longitude) dos ônibus e registro de data e hora, presentes no DST-A. No entanto, é preciso retirar os ônibus que não estão na região de interesse (Município do Rio de Janeiro) e os que estão parados em garagens e terminais, porque podem induzir a geração de *clusters* nesses locais e poluir a base que será analisada. Para isso, são usados os atributos *NEIGHBORHOOD*, *PARKING* e *TERMINAL*, presentes no DST-B.

Ainda antes de executar o DBSCAN é feito mais um filtro nos dados, pois, como se objetiva encontrar os ônibus inseridos em pontos de alagamento intransponíveis, as observações são filtradas de modo a retornarem apenas os ônibus que estão parados, ou seja, com os atributos *SPEED* e *DISTANCE*, presentes no DST-D, iguais a zero.

VI.2 Partição Horária

Para evitar que sejam gerados *clusters* em observações que ocorreram em horários muito distantes, os dados são separados em partições horárias de 30 minutos. Essa partição é definida de modo a ficar compatível com os registros de volume de chuva que são utilizados *a posteriori* e que possuem partição horária de 15 minutos, conforme descrito na Seção IV.1.

VI.3 Clusterização

A clusterização tem como objetivo identificar os *clusters* de registros dos ônibus em locais com um volume de chuva acumulado nas últimas horas relevante. Foi utilizada a implementação do DBSCAN da biblioteca *sklearn.cluster* em Python para identificação dos *clusters*.

Os parâmetros de *epsilon* e *MinPts* escolhidos foram 20 e 10, respectivamente. Dessa forma, significa que o raio de alcance do *cluster* é 20 metros e o *MinPts* são 10 observações.

VI.4 Agrupamento dos *clusters*

Nessa etapa, são retiradas as observações que não estão contidas em nenhum *cluster*, ou seja, são ruídos. Os dados que se mantém são agrupados de modo que o conjunto de dados trabalhado passe a conter apenas as informações dos *clusters*, de partição horária, o ID e o atributo de geometria do *cluster*. Esse atributo de geometria informa a posição (latitude e longitude) de todos os pontos que estão inseridos naquele *cluster*.

VI.5 Análise do Volume de Chuva

Para a validação se o *cluster* está inserido em uma área de alagamento, verifica-se se volume de chuva na última hora na zona pluviométrica que o *cluster* está inserido é maior que 5 milímetros. Esse valor é o definido pelo AlertaRio como limite de uma chuva fraca, acima disso é considerada como chuva moderada, conforme Tabela VI.1. Nessa etapa são usados os dados apresentados na Seção IV.1.

Categoria	**Descrição**
Chuva fraca	Precipitação cuja intensidade é menor do que 5,0 mm/h.
Chuva moderada	Precipitação cuja intensidade está compreendida entre 5,0 e 25 mm/h.
Chuva forte	Precipitação cuja intensidade está compreendida entre 25,1 e 50 mm/h.
Chuva muito forte	Precipitação cuja intensidade é maior do que 50,0 mm/h.
Pancadas de chuva	Precipitação intensa ocorrida em curto período de tempo e espacialmente restrita

Tabela VI.1: Categorias de chuva pelo AlertaRio

VI.6 *Cluster* por logradouros

Após a identificação dos *clusters* inseridos em regiões de chuva moderada a muito forte, é feito o cruzamento do conjunto de dados gerado com o *dataset* de logradouros disponibilizado pela Prefeitura, por meio do Data.Rio, de modo que seja retornado para cada *cluster* o nome da rua em que ele está inserido.

VI.7 Análise comparativa com pontos de alagamento históricos

A última etapa consiste em uma análise comparativa entre os agrupamentos espaciais identificados a partir dos dados de mobilidade da frota de ônibus e os registros históricos de alagamentos no município do Rio de Janeiro. Essa análise tem como propósito avaliar a correspondência espacial entre os *clusters* que indicam interrupções na operação dos ônibus e os pontos historicamente reconhecidos como recorrentes em eventos de alagamento.

São definidas três categorias para análise comparativa: Totalmente compatível, parcialmente compatível e não compatível, descritas na Tabela VI.2. A definição dessas categorias possibilita obter resultados quantitativos para a análise qualitativa. Dessa forma é possível definir métricas e avaliar a eficácia da metodologia aplicada.

No estudo proposto, essa etapa é fundamental pois ela valida, ou não, a metodologia proposta e os parâmetros adotados para geração dos *clusters*. Caso o resultado obtido não esteja satisfatório, deve-se retornar para a etapa de Clusterização e revisar os parâmetros.

Categoria	**Descrição**
Totalmente Compatível	Atributo posicional do *cluster* sobreposto precisamente ao local citado na reportagem
Parcialmente Compatível	Atributo posicional do *cluster* sobreposto próximo ao local citado na reportagem
Não compatível	Não foi possível associar o *cluster* ao local da reportagem

Tabela VI.2: Resultado da Análise Comparativa

VI.8 Resultados Análise Comparativa

Para realizar a análise comparativa e verificar se a metodologia proposta retornaria pontos de alagamentos históricos, foi utilizada uma reportagem veiculada na TV Record no dia 22/02/2023, que pode ser acessada através do link `https://bit.ly/3BKxRse`. Na reportagem é possível observar dois veículos presos em um alagamento e que estão parcialmente submersos no bairro de Bonsucesso, conforme Figura VI.1 .

Dessa forma, como o objetivo desta usabilidade é encontrar locais em que se formam alagamentos intransponíveis, o dia 22/02/2023 foi definido como o objeto de estudo e, visto que os códigos de identificação dos ônibus estão visíveis, encontrar o *cluster* em que esses veículos estão inseridos se tornou uma meta.

Foi possível encontrar os registros desses ônibus no conjunto de dados gerado pelo *pipeline* e foi observado que os registros estão condizentes tanto em localização quanto em permanência de tempo parado. Foi observado, também, que o volume de chuva retornado na base estava condizente



Figura VI.1: Ônibus presos em ponto de alagamento

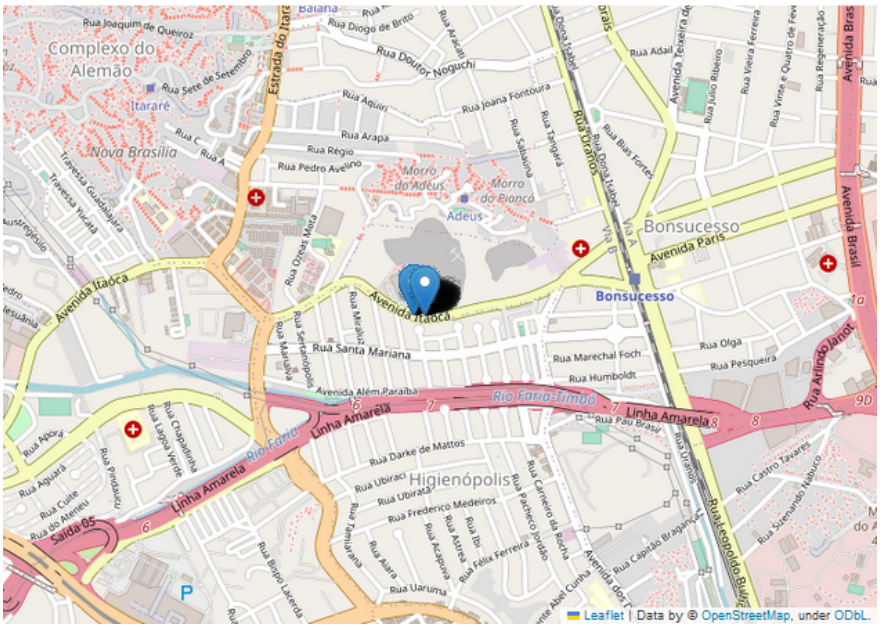


Figura VI.2: Registro GPS dos ônibus presos no ponto de alagamento da reportagem

com o momento em que eles se encontraram presos no alagamento, conforme Figura VI.2.

Uma vez identificado que os registros dos ônibus estão inseridos no conjunto de dados, foi analisado se essas observações geraram um *cluster* na localidade. Essa busca pôde ser feita por meio do nome da rua veiculada na reportagem e foi observado que o *cluster* havia sido gerado, conforme a Figura VI.3.

A partir da identificação desse ponto de alagamento e de uma primeira validação da metodologia, foi realizada outra análise comparativa. Dessa vez baseada em uma reportagem veiculada no site do Jornal O Globo, no dia 31/03/2022, que pode ser acessada através do link `http://bit.ly/4lki8lW`, onde são listados diversos pontos de bolsões no Município do Rio de Janeiro. Para cada logradouro apresentado na reportagem, foi verificado se havia algum *cluster* retornado pela metodologia.

Os resultados estão apresentados na Tabela VI.3. Definido como resultado positivo os *clusters*

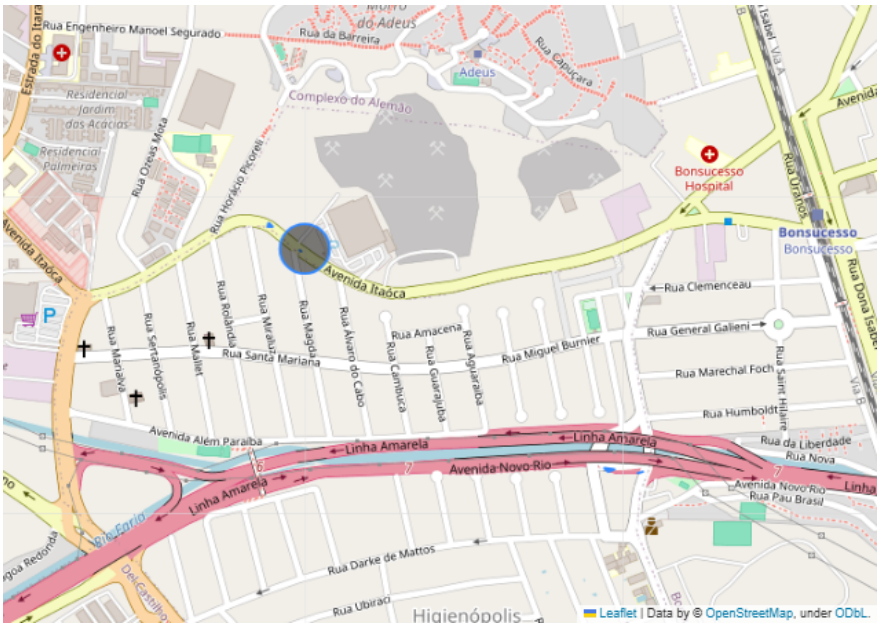


Figura VI.3: *Cluster* formado pelas observações dos ônibus identificados no ponto de alagamento da reportagem

que sejam ao menos parcialmente compatível, a metodologia obteve uma taxa de positivos reais de 43,50%. Vale ressaltar que a matéria utilizada como referência, apresenta pontos de bolsões d'água, que não necessariamente são alagamentos de muita relevância em que os veículos não conseguem ultrapassar. Dessa forma, entende-se que a metodologia foi eficaz para a identificação de alagamentos intransponíveis.

Categoria	**Quantidade**
Totalmente Compatível	12
Parcialmente Compatível	8
Não compatível	26

Tabela VI.3: Resultado da Análise Comparativa

Capítulo VII Conclusão

Nesse capítulo são apresentadas as considerações finais para o trabalho apresentado e suas oportunidades de trabalhos futuros, bem como a sua contribuição para a academia.

VII.1 Considerações finais

Nesse trabalho é apresentada a construção de uma base de dados integrada e enriquecida para análise da mobilidade urbana na cidade do Rio de Janeiro. A metodologia desenvolvida foi capaz de combinar dados de GPS dos ônibus da cidade, com informações posicionais, pluviométricas e cálculos de emissões de gases, possibilitando análises para diversas áreas.

A proposta incluiu o desenvolvimento de um pipeline robusto para extração, limpeza, enriquecimento e agregação dos dados, resultando em cinco conjuntos DSTs com diferentes focos: localização, administração urbana, clima, deslocamento e emissões. A qualidade e granularidade da base gerada possibilitam diversos tipos de análise, como estudos sobre emissão de poluentes, identificação de gargalos operacionais e monitoramento de zonas críticas da cidade.

A análise de usabilidade demonstrou que a metodologia proposta é eficaz na detecção de eventos anômalos, como alagamentos, sobretudo quando validada por registros externos e históricos. O cruzamento com reportagens jornalísticas mostrou que os *clusters* gerados correspondiam, em sua maioria, a pontos de alagamento reconhecidos, validando empiricamente a abordagem.

Como principal contribuição, o trabalho oferece uma base pública, atualizada e integrada de mobilidade urbana, promovendo avanços no uso de dados abertos para fins de planejamento urbano e tomada de decisão informada por parte do poder público.

VII.2 Trabalhos Futuros

A elaboração e disponibilização da base de dados fruto deste trabalho possibilita uma gama de análises para o planejamento urbano, desde o âmbito ambiental com o volume de gases poluentes emitidos, até os efetivos pontos de engarrafamento nos bairros e regiões pluviométricas.

Podemos citar como oportunidade de trabalho futuro, também, a utilização de uma maior base de informações históricas sobre os pontos de alagamento e bolsões d'água na cidade, tais como redes sociais oficiais das autoridades municipais e fontes confiáveis de notícias.

A taxa de sucesso apresentada pela metodologia apresentada na seção VI. incentiva a aplicação de uma possível adaptação em dados posicionais com características diferentes para analisar a mobilidade urbana. Esses dados posicionais podem ser, por exemplo, dados das câmeras de segurança da cidade ou de radares de trânsito, que mapeiam o volume de veículos que se movimentaram naquele período na região mapeada por aquele instrumento, cruzando com o *dataset* de volume pluviométrico.

Apêndice A Zonas Pluviométricas do município do Rio de Janeiro

OBJECTID	Name	Est
1	Região 1	Vidigal
2	Região 2	Urca
3	Região 3	Rocinha
4	Região 4	Tijuca
5	Região 5	Santa Teresa
6	Região 6	Copacabana
7	Região 7	Grajaú
8	Região 8	Ilha do Governador
9	Região 9	Penha
10	Região 10	Madureira
11	Região 11	Irajá
12	Região 12	Bangu
13	Região 13	Piedade
14	Região 14	Jacarepaguá/Tanque
15	Região 15	Saúde
16	Região 16	Jardim Botânico
17	Região 17	Barra/Itanhangá
18	Região 18	Jacarepaguá/Cidade de Deus
19	Região 19	Barra/Rio Centro
20	Região 20	Guaratiba
21	Região 21	Estrada Grajaú/Jacarepaguá
22	Região 22	Santa Cruz
23	Região 23	Grande Méier
24	Região 24	Anchieta
25	Região 25	Grota Funda
26	Região 26	Campo Grande
27	Região 27	Sepetiba
28	Região 28	Alto da Boa Vista
29	Região 29	Av. Brasil/Mendanha
30	Região 30	Recreio dos Bandeirantes
31	Região 31	Laranjeiras
32	Região 32	São Cristóvão
33	Região 33	Tijuca/Muda

Tabela A.1: Zonas Pluviométricas e seus Códigos

Apêndice B Bairros do município do Rio de Janeiro

CODRA	Região Administrativa	CODBNUM	Bairros (ordenados)
1	PORTUARIA	002, 003, 001, 004	Gamboa, Santo Cristo, Saúde, Caju
2	CENTRO	005, 161	Centro, Lapa
3	RIO COMPRIDO	008, 007, 009, 006	Cidade Nova, Rio Comprido, Estácio, Catumbi
4	BOTAFOGO	016, 018, 015, 017, 019, 020, 022, 021	Glória, Catete, Flamengo, Laranjeiras, Cosme Velho, Botafogo, Urca, Humaitá
5	COPACABANA	024, 023	Copacabana, Leme
6	LAGOA	028, 027, 029, 026, 025, 031, 030	Jardim Botânico, Lagoa, Gávea, Leblon, Ipanema, São Conrado, Vidigal
7	SAO CRISTOVAO	012, 010, 158, 011	Benfica, São Cristóvão, Vasco da Gama, Mangueira
8	TIJUCA	032, 033, 034	Praça da Bandeira, Tijuca, Alto da Boa Vista
9	VILA ISABEL	036, 035, 038, 037	Vila Isabel, Maracanã, Grajaú, Andaraí
10	RAMOS	042, 041, 040, 039	Olaria, Ramos, Bonsucesso, Manguinhos
11	PENHA	044, 045, 043	Penha Circular, Brás de Pina, Penha
12	INHAUMA	055, 054, 056, 050, 053, 052	Engenho da Rainha, Inhaúma, Tomás Coelho, Higienópolis, Del Castilho, Maria da Graça
13	MEIER	071, 069, 070, 065, 066, 064, 051, 068, 058, 063, 060, 059, 057, 061, 067, 062	Pilares, Piedade, Abolição, Cachambi, Engenho de Dentro, Todos os Santos, Jacaré, Encantado, Rocha, Méier, Sampaio, Riachuelo, São Francisco Xavier, Engenho Novo, Água Santa, Lins de Vasconcelos
14	IRAJA	076, 075, 077, 074, 072, 073	Irajá, Vista Alegre, Colégio, Vila da Penha, Vila Kosmos, Vicente de Carvalho
15	MADUREIRA	087, 086, 090, 084, 089, 085, 083, 088, 080, 081, 082, 079, 078	Honório Gurgel, Rocha Miranda, Marechal Hermes, Vaz Lobo, Bento Ribeiro, Turiaçú, Madureira, Osvaldo Cruz, Cavalcanti, Engenheiro Leal, Cascadura, Quintino Bocaiúva, Campinho
16	JACAREPA-GUA	125, 124, 123, 122, 120, 121, 119, 116, 115, 117	Vila Valqueire, Praça Seca, Tanque, Taquara, Freguesia (Jacarepaguá), Pechincha, Curicica, Anil, Jacarepaguá, Gardênia Azul
17	BANGU	140, 142, 141, 160, 162, 163	Padre Miguel, Senador Camará, Bangu, Gericinó, Vila Kennedy, Jabour
18	CAMPO GRANDE	143, 145, 147, 146, 144	Santíssimo, Senador Vasconcelos, Cosmos, Inhoaíba, Campo Grande
19	SANTA CRUZ	148, 149, 150	Paciência, Santa Cruz, Sepetiba
20	ILHA DO GOVERNADOR	098, 097, 104, 101, 103, 102, 096, 100, 099, 095, 093, 094, 092, 091, 105	Freguesia (Ilha), Bancários, Galeão, Tauá, Portuguesa, Moneró, Cocotá, Jardim Carioca, Jardim Guanabara, Praia da Bandeira, Cacuia, Pitangueiras, Zumbi, Ribeira, Cidade Universitária
21	PAQUETA	013	Paquetá
22	ANCHIETA	107, 106, 108, 109	Anchieta, Guadalupe, Parque Anchieta, Ricardo de Albuquerque
23	SANTA TEREZA	014	Santa Teresa
24	BARRA DA TIJUCA	131, 129, 130, 127, 128, 132, 126, 133	Vargem Grande, Camorim, Vargem Pequena, Itanhangá, Barra da Tijuca, Recreio dos Bandeirantes, Joá, Grumari
25	PAVUNA	114, 159, 111, 113, 110, 112	Pavuna, Parque Colúmbia, Acari, Costa Barros, Coelho Neto, Barros Filho
26	GUARATIBA	151, 153, 152, 164	Guaratiba, Pedra de Guaratiba, Barra de Guaratiba, Ilha de Guaratiba
27	ROCINHA	154	Rocinha
28	JACAREZINHO	155	Jacarezinho
29	COMPLEXO DO ALEMÃO	156	Complexo do Alemão
30	COMPLEXO DA MARE	157	Maré
31	VIGARIO GERAL	048, 049, 046, 047	Vigário Geral, Jardim América, Cordovil, Parada de Lucas
33	REALENGO	135, 139, 138, 136, 137, 134	Vila Militar, Realengo, Magalhães Bastos, Campo dos Afonsos, Jardim Sulacap, Deodoro
34	CIDADE DE DEUS	118	Cidade de Deus

Tabela B.1: Zonas Administrativas e Bairros do Rio de Janeiro

Referências Bibliográficas

Haibo Zhai, H. Christopher Frey, and Nagui M. Rouphail. A vehicle-specific power approach to speed- and facility-specific emissions estimates for diesel transit buses. *Environmental Science and Technology*, 42(21):7985 – 7991, 2008. doi: 10.1021/es800208d. , 15, 16

Luis Bettencourt and Geoffrey West. A unified theory of urban living. *Nature*, 467(7318):912–913, Oct 2010. ISSN 1476-4687. doi: 10.1038/467912a. URL `https://doi.org/10.1038/467912a`. 1

Aris Christodoulou, Lewis Dijkstra, Panayotis Christidis, Paolo Bolsi, and Hugo Poelman. A fine resolution dataset of accessibility under different traffic conditions in european cities. *Scientific Data*, 7(1):279, Aug 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00619-7. URL `https://doi.org/10.1038/s41597-020-00619-7`. 1

Qiuyang Huang, Yongjian Yang, Zhilu Yuan, Hongfei Jia, Liping Huang, and Zhanwei Du. The temporal geographically-explicit network of public transport in changchun city, northeast china. *Scientific Data*, 6(1):190026, Feb 2019. ISSN 2052-4463. doi: 10.1038/sdata.2019.26. URL `https://doi.org/10.1038/sdata.2019.26`. 1

Vincent Verbavatz and Marc Barthelemy. Critical factors for mitigating car traffic in cities. *PLOS ONE*, 14(7):1–16, 07 2019. doi: 10.1371/journal.pone.0219559. URL `https://doi.org/10.1371/journal.pone.0219559`. 1

Vincent Verbavatz and Marc Barthelemy. Access to mass rapid transit in oecd urban areas. *Scientific Data*, 7(1):301, Sep 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00639-3. URL `https://doi.org/10.1038/s41597-020-00639-3`. 1

Rémi Louf and Marc Barthelemy. How congestion shapes cities: from mobility patterns to scaling. *Scientific Reports*, 4(1):5561, Jul 2014. ISSN 2045-2322. doi: 10.1038/srep05561. URL `https://doi.org/10.1038/srep05561`. 1

Aris Christodoulou and Panayotis Christidis. Measuring congestion in european cities. *Publications Office of the European Union, Luxembourg, URL: https://publications. jrc. ec. europa. eu/repository/handle/JRC118448*, 2020. 1

Ana Beatriz Cruz, João Ferreira, Diego Carvalho, Eduardo Mendes, Esther Pacitti, Rafaelli Coutinho, Fábio Porto, and Eduardo Ogasawara. Detecçao de Anomalias Frequentes no Transporte Rodoviario Urbano. In *SBBD: Simpósio Brasileiro de Banco de Dados*, pages 271–276, Rio de Janeiro, Brazil, August 2018. SBC, SBC. URL `https://hal-lirmm.ccsd.cnrs.fr/lirmm-01868597`. 2, 5

Heraldo Borges, Murillo Dutra, Amin Bazaz, Rafaelli Coutinho, Fábio Perosi, Fábio Porto, Florent Masseglia, Esther Pacitti, and Eduardo Ogasawara. Spatial-Time Motifs Discovery. *Intelligent Data Analysis*, October 2020. URL `https://hal-lirmm.ccsd.cnrs.fr/lirmm-02984969`. 4

Mingyang Zhang, Tong Li, Yue Yu, Yong Li, Pan Hui, and Yu Zheng. Urban anomaly analytics: Description, detection and prediction. *IEEE Transactions on Big Data*, pages 1–1, 2020. doi: 10.1109/TBDATA.2020.2991008. 5

Ramez Elmasri and Shamkant B. Navathe. *Fundamentals of Database Systems*. Pearson, 6 edition, 2011. 5

C. J. Date. *An Introduction to Database Systems*. Pearson Education, 8 edition, 2004. 5

Xiangfu Kong, Jiawen Yang, Jiandong Qiu, Qin Zhang, Xunlai Chen, Mingjie Wang, and Shan Jiang. Post-event flood mapping for road networks using taxi gps data. *Journal of Flood Risk Management*, 15(2):e12799, 2022. 6

Riches S Bacero and Alexis M Fillone. Impact of flooding on truck movement in metro manila, philippines. *Asian Transport Studies*, 9:100113, 2023. 6

Weipeng Jing, Jian Kang, and Meiling Liu. Mining taxi trajectories for most suitable stations of sharing bikes to ease traffic congestion. *IET Intelligent Transport Systems*, 12(7):586–593, 2018. 6

Sharmili Banik, Darcy M Bullock, and Lelitha Vanajakshi. Corridor level mobility analysis using gps data. *International Journal of Intelligent Transportation Systems Research*, 18:204–218, 2020. 6

Haoze Wang and Yutong Si. Detection of traffic abnormity based on clustering analysis of taxi gps data. In *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, pages 219–224, 2019. 7

Qiuping Li, Haowen Luo, and Xuechen Luan. Multistage impacts of the heavy rain process on the travel speeds of urban roads. *ISPRS International Journal of Geo-Information*, 10(8):557, 2021. 7

J. Koupal, M. Cumberworth, H. Michaels, M. Beardsley, and D. Brzezinski. Draft design and implementation plan for EPA's multi-scale motor vehicle and equipment emission systems (MOVES). Technical report, https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P1000527.PDF, 2002. 15

Nigel N. Clark, Justin M. Kern, Christopher M. Atkinson, and Ralph D. Nine. Factors affecting heavy-duty diesel vehicle emissions. *Journal of the Air and Waste Management Association*, 52 (1):84 – 94, 2002. doi: 10.1080/10473289.2002.10470755. 15

Tom G. Farr, Paul A. Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, David Seal, Scott Shaffer, Joanne Shimada, Jeffrey Umland, Marian Werner, Michael Oskin, Douglas Burbank, and Douglas Alsdorf. The shuttle radar topography mission. *Reviews of Geophysics*, 45(2), May 2007. doi: 10.1029/2005rg000183. URL `https://doi.org/10.1029/2005rg000183`. 17