BRBus - construindo um *dataset* para monitoramento geoespacial dos ônibus de cidades brasileiras

Ruan T. Melo[1], Felipe F. Vasconcelos[1], Rafael Luciano L. Silva[1], Pedro Victor Santos[1], Vinicius T. Ramos[1], Fábio J. Coutinho[1]

[1]Instituto de Computação – Universidade Federal Alagoas (UFAL)

{rtm,ffv,rlls,pvafs,vtpr,fabio}@ic.ufal.br

Abstract. *In Brazil, according to ANTP data, in cities with more than 60,000 inhabitants, 85% of trips made by public transport take place using buses. Considering the problem of Urban Mobility, geospatial data produced by different devices (e.g. buses, personal cars, traffic lights, radars) offer great analytical potential, being relevant for making decisions that impact the quality of life in smart cities. In this work, we describe the process of collecting, standardizing and enriching BRBus – a dataset with information from the geospatial monitoring of bus traffic in four Brazilian cities. BRBus is available in an open format and covers data collected between 12/06/2023 and 20/06/2023.*

Resumo. *No Brasil, segundo dados da ANTP, em cidades com mais de 60 mil habitantes, 85% das viagens realizadas por transporte público ocorrem utilizando ônibus. Considerando o problema da Mobilidade Urbana, dados geoespaciais produzidos por diferentes dispositivos (e.g. ônibus, carros pessoais, semáforos, radares) oferecem um grande potencial analítico, sendo relevante para a tomada de decisões que impactam a qualidade de vida em cidades inteligentes. Neste trabalho, descrevemos o processo de coleta, padronização e enriquecimento do BRBus – um dataset com informações do monitoramento geoespacial do tráfego de ônibus de quatro cidades brasileiras. BRBus está disponível em formato aberto e abrange dados coletados entre 12/06/2023 e 20/06/2023.*

1. Introdução

De acordo com o portal Statista, mais de 75 bilhões de dispositivos IoT estarão conectados até o ano de 2025 [Statista 2016]. Essa marca reflete o avanço no desenvolvimento das cidades inteligentes, fato que eleva significativamente o volume de dados produzidos nos últimos anos. As soluções implementadas neste contexto, frequentemente, voltam-se para áreas de interesse dos gestores das cidades tais como administração, saúde, meteorologia, mobilidade urbana, segurança, poluição, entre outros. Em se tratando da Mobilidade Urbana, dados geospaciais produzidos por diferentes dispositivos, tais como veículos, sensores, semáforos, radares e outros, oferecem um grande potencial analítico, sendo relevante para a tomada de decisões que impactam a qualidade de vida em cidades inteligentes [Alablani and Alenazi 2020]. Técnicas para gerenciamento do transporte público [TAO 2013], detecção de congestionamento [Jose and Mitra 2018] e busca por rotas mais eficientes [Braz et al. 2018] são alguns exemplos de problemas enfrentados em grandes centros urbanos que podem tirar proveito da análise de dados geoespaciais.

A Mobilidade Urbana, em grandes cidades, pode ser avaliada a partir de indicadores como a proporção de automóveis por habitantes, a idade média da frota de veículos, a relação de ônibus *versus* automóveis, a presença de ciclovias, entre outros. No Brasil, segundo dados da ANTP, em cidades com mais de 60 mil habitantes, 85% das viagens realizadas por transporte público ocorrem utilizando ônibus [de Transporte Público 2018]. Este trabalho descreve o processo de desenvolvimento do BRBus – um dataset acerca de dados de monitoramento geoespacial do transporte público por ônibus de quatro cidades brasileiras[1] (São Paulo-SP, Curitiba-PR, Brasília-DF e Rio de Janeiro-RJ), classificadas entre as sete primeiras colocadas do país, de acordo com a edição 2022 do Ranking Connected Smart Cities [CSC 2022].

As cidades selecionadas dispõem de frotas que variam de cerca de 4,6 a 11 mil ônibus e fornecem uma API específica para coleta de dados que incluem diferentes informações, tais como posição geográfica, letreiro do ônibus, tipo do veículo, entre outros. A construção do *dataset* abrange as seguintes etapas: coleta de dados, identificação de problemas, padronização, limpeza e enriquecimento de dados. O objetivo principal deste trabalho consiste em fomentar a colaboração e pesquisa sobre tópicos acerca da mobilidade urbana, engenharia de tráfego, urbanismo inteligente, segurança no trânsito e outros. O *dataset* BRBus permite que gestores e demais interessados possam realizar análises exploratórias dos sistemas de transporte público por ônibus de grandes cidades brasileiras, possibilitando análises de comportamento do sistema e análises comparativas entre as cidades. O *dataset* é disponibilizado gratuitamente em formato aberto, podendo ser lido em qualquer plataforma ou ferramenta de análise.

Aplicações para os dados de transporte público de cidades brasileiras, como os do BRBus, são recorrentes na literatura. [Queiroz et al. 2019] utilizou dados de movimentação da frota de ônibus do Rio de Janeiro e de Curitiba para analisar a qualidade dos dados GTFS (General Transit Feed Specification) disponibilizados. [Arbex and da Cunha 2016] utilizaram dados geoespaciais da cidade de São Paulo para analisar possíveis mudanças na velocidade dos ônibus após a instalação de faixas exclusivas. [Larsen et al. 2020] utilizaram dados de geolocalização da frota de ônibus de São Paulo, em conjunto com dados de trafégo e informações climáticas, para treinar uma rede neural que realizasse a predição do tempo de viagem dos ônibus.

Este documento encontra-se organizado da seguinte maneira: A seção 2 discute os trabalhos relacionados à temática abordada em nosso trabalho; Na seção 3, são descritas as fontes de dados e as dificuldades encontradas na utilização; A seção 4 relata o processo de construção do BRBus, abordando as etapas de extração, pré-processamento, padronização e enriquecimento, além do dicionário de dados; A seção 5 apresenta algumas aplicações e experimentos realizados nos dados do BRBus, como mapas de calor e uma análise quantitativa da frota ativa; A seção 6 apresenta as considerações finais, trabalhos futuros e as formas de acesso ao *dataset*.

2. Trabalhos Relacionados

Alguns trabalhos encontrados na literatura propõem a construção e/ou manipulação de *datasets* contendo dados geoespaciais referentes ao monitoramento de tráfego de ônibus

[1]Os dados também incluem ônibus que circulam em cidades vizinhas que integram a região metropolitana

[Cruz Junior 2020], [Queiroz et al. 2019], [da Cruz et al. 2016]. Entretanto, esses trabalhos não têm como finalidade principal a construção de um *dataset* para uso de terceiros, conforme descrito a seguir.

[Cruz Junior 2020] construiu um *dataset* e realizou experimentos com dados de geolocalização e bilhetagem eletrônica de ônibus municipais de Curitiba. Os dados utilizados foram extraídos de APIs disponibilizadas pela cidade de Curitiba. O estudo utilizou a linguagem R para realizar transformações e análises nos dados obtidos, apresentando uma análise espacial das origens e destinos finais em horários de picos e a detecção de outliers em viagens. Todavia, diferentemente do nosso trabalho, o autor não disponibiliza publicamente o *dataset* gerado.

[da Cruz et al. 2016] desenvolveram uma aplicação móvel para disponibilizar informações em tempo real sobre as linhas de ônibus da cidade do Rio de Janeiro. Os dados dos trajetos das linhas e das localizações atualizadas dos ônibus foram coletados de *datasets* fornecidos pela Prefeitura do Rio de Janeiro, disponíveis no website data.rio[2]. Os autores implementaram um processo que inclui extração com crawlers, limpeza e transformações nos dados, persistindo-os em servidores NoSQL (MongoDB e Cassandra) e, posteriormente, disponibilizando-os através de uma API para uso exclusivo da aplicação. Ou seja, diferentemente de nossa proposta, não há publicação do *dataset* construído nesse trabalho.

[Queiroz et al. 2019] produziram uma análise de conformidade entre os dados históricos das rotas em GTFS e os dados em tempo real das trajetórias de ônibus de três cidades brasileiras: Curitiba, Rio de Janeiro e uma outra cidade que não foi revelada. A partir de dados disponibilizados pelas agências de transporte público das cidades, foi aplicado o método de map matching, utilizando dados do GPS para mapear as rotas dos ônibus a fim de medir a similaridade entre os dados do GPS e do GTFS. As análises destacaram problemas de inconsistência nos dados e ofereceram uma abordagem para melhorar a identificação das rotas seguidas pelos ônibus. Portanto, identifica-se a necessidade real, por parte desses autores, de *datasets* já preparados para uso tal como o BRBus, o que evitaria a realização de processos individuais de coleta e pré-processamento.

Os trabalhos discutidos anteriormente evidenciam o interesse da comunidade científica pelo monitoramento geoespacial de veículos, com ênfase nos ônibus municipais. BRBus diferencia-se por abranger dados de um maior número de cidades brasileiras, realizando não apenas o processo de coleta e limpeza, mas promovendo também o enriquecimento do *dataset* com informações de outras fontes. Essa abordagem abrangente, integrada e compartilhada oferece uma oportunidade de realizar análises exploratórias intercidades, fornecendo insights valiosos para os tomadores de decisão.

3. Dados do tráfego de ônibus de cidades brasileiras

A etapa inicial de construção do *dataset* consiste na coleta de dados a partir de requisições a quatro diferentes APIs, as quais são disponibilizadas pelos departamentos ou empresas responsáveis pela mobilidade urbana das respectivas cidades, conforme descrito a seguir:

- São Paulo Transporte S/A - SPTrans[3] (São Paulo).

[2]https://www.data.rio

[3]https://www.sptrans.com.br

- Empresa Municipal de Informática - IplanRio[4] (Rio de Janeiro).
- Secretaria de Transporte e Mobilidade de Brasília[5] (Brasília).
- Urbanização de Curitiba S/A - URBS[6] (Curitiba).

A coleta também inclui a obtenção de dados complementares através do acesso à base de dados do Instituto Brasileiro de Geografia e Estatística (IBGE) e da plataforma Moovit[7], que dispõe de informações sobre transportes públicos de diferentes modais. Na Tabela 1, pode-se visualizar as informações disponibilizadas por cada cidade.

	Cidades			
Informações Disponíveis	São Paulo	Rio de Janeiro	Brasília	Curitiba
Coordenadas	X	X	X	X
Identificador do ônibus	X	X	X	X
Velocidade instantânea		X	X	
Horário de atualização	X	X	X	X
Código da linha	X	X	X	X
Situação de atraso				X
Sentido da rota	X		X	X
Tipo do veículo			X	
Acessibilidade a deficientes	X			
Informação do letreiro	X			
Quantidade de veículos na linha	X			
Número de ciclos sem atualização				X

Tabela 1. Informações disponibilizadas pelas cidades analisadas

3.1. Problemas identificados nos dados originais

Considerando as fontes de dados utilizadas para a construção do *dataset*, foram identificadas divergências no dicionário de dados de cada operadora, visto que cada uma registrava os dados em esquema próprio. Além disso, foram encontradas outras questões relacionadas à ausência de dados, esses problemas são descritos em detalhes a seguir.

- Nos dados de Curitiba, o campo “REFRESH”, referente à última atualização da posição do ônibus, consta apenas com o horário no formato ”HH:mm”, não sendo possível identificar a data.
- Cada fonte utiliza um dicionário de dados próprio.
- As fontes de cada cidade utilizam diferentes formatos para representar o tempo em seus respectivos campos, como por exemplo, os dados de São Paulo utilizam o formato “YYYY-MM-DD HH:mm:ssZ”, enquanto os do Rio de Janeiro utilizam “MM-DD-YYYY HH:mm:ss”.
- A API de Curitiba apresenta todos os campos no formato String, fato que diverge das outras APIs, que atribuem tipagem aos campos de acordo com os valores.
- Ausência de dados referentes às operadoras e aos valores das tarifas das linhas de ônibus.
- As fontes de dados podem reportar ônibus fora de operação ou com defeitos no dispositivo de geolocalização.

[4]https://iplanrio.prefeitura.rio
[5]https://semob.df.gov.br
[6]https://www.urbs.curitiba.pr.gov.br
[7]https://moovitapp.com

4. Construção do *dataset* BRBus

Durante a construção do *dataset* foram realizadas as seguintes etapas: (i) extração - obtenção dos dados das APIs e do Moovit; (ii) limpeza - que consiste na remoção de arquivos e ônibus inoperantes; (iii) padronização - que compreende a criação de um dicionário de dados comum a todas as fontes de dados e na formatação padrão de campos; (iv) enriquecimento - adição de novas informações de diferentes fontes, como dados do IBGE e dados do Moovit. A Figura 1 demonstra o fluxo das atividades e as ferramentas utilizadas na construção do *dataset*.

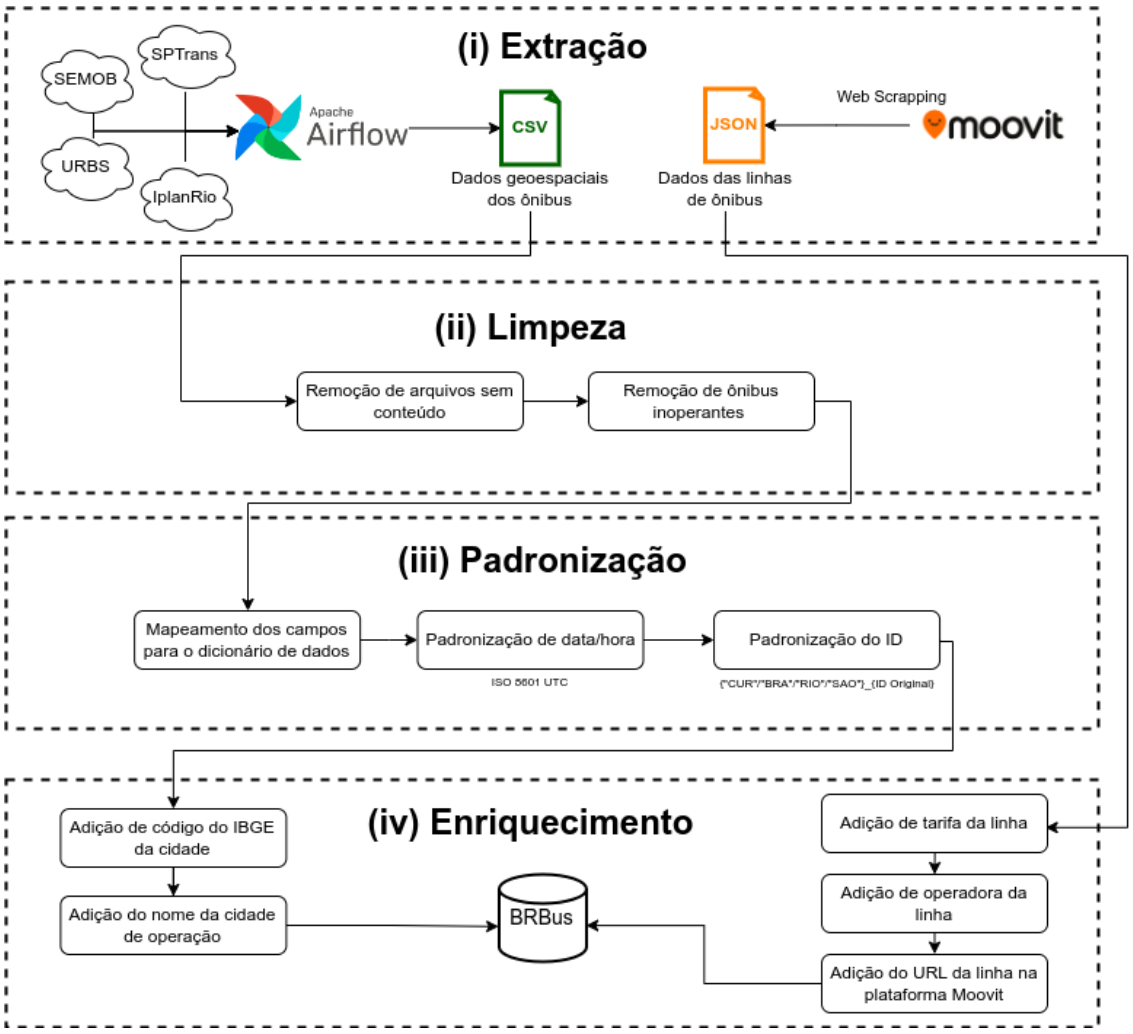


Figura 1. Diagrama do fluxo de construção do BRBus

Essas etapas são necessárias visto que cada cidade apresenta seu próprio esquema, fonte e formato de dados. Em decorrência disso, uma abordagem de curadoria manual se faz necessária a fim de ajustar os dados originais ao padrão criado pelo BRBus.

Devido à popularidade e à facilidade de uso, todas as etapas do processo foram implementadas utilizando Python e bibliotecas como: Beautiful Soup [8] e Selenium [9].

[8]https://www.crummy.com/software/BeautifulSoup/
[9]https://www.selenium.dev/

4.1. Extração

O processo de extração inicia-se a partir da obtenção dos dados geoespaciais dos ônibus, através de requisições às APIs, realizadas a cada 60 segundos, por meio de tarefas agendadas pela ferramenta Apache Airflow. Os dados foram coletados durante o período de 12/06/2023 até 20/06/2023 e armazenados localmente no formato *Comma-Separated Values* (CSV), totalizando 84GB, sendo: 45GB para São Paulo; 28GB para Brasília; 7GB para Rio de Janeiro; 4GB para Curitiba.

Além do processo de coleta das informações das APIs, também foi realizado um processo de extração de dados acerca das linhas de ônibus através de web scraping na plataforma Moovit, coletando dados a serem usados conforme descrito na Seção 4.3.

Após a coleta dos dados ter sido realizada, iniciou-se a fase de limpeza das respostas das APIs, atuando nos casos descritos abaixo:

- Durante o processo de coleta foram identificados e removidos arquivos que não apresentavam conteúdo, seja por erros no servidor de API ou por instabilidades de conexão.
- Nas respostas das APIs, ônibus que possuíam uma diferença de 5 minutos entre horário da requisição e o horário do último ping de atualização dos dados pela empresa responsável.
- Foi identificado que as respostas da API de Curitiba apresentavam dados vazios quando o campo original "CODIGOLINHA" tinha valor "REC", logo, foram removidas do arquivo da requisição.
- Ausência de valor no campo "LINHA" na API de Brasília indicava que o ônibus não estava em operação. Sendo assim, os dados destes veículo foram removidos.

4.2. Padronização

Esta etapa compreende a criação de um dicionário de dados por meio da identificação dos diferentes esquemas utilizados por cada cidade para representar campos similares e a conversão dos dados para um formato padrão, com o objetivo de manter a homogeneidade dos dados. As seguintes informações foram utilizadas no dicionário de dados descrito na Seção 4.4 e afetadas pela etapa de padronização:

- **Horário de atualização da informação do ônibus:**
  - Cada API utilizava um formato para representar o horário de atualização da informação do ônibus. Todos os dados foram convertidos para o formato ISO 8601 com fuso horário UTC.
- **Sentido de operação do ônibus**
  - As APIs de São Paulo e Curitiba informam o sentido de operação da linha, porém divergem na representação do dado. Estes campos foram convertidos e padronizados para números inteiros conforme descrito na tabela 2, representando os sentidos de ida e volta.
- **Identificadores dos ônibus**
  - Os identificadores dos ônibus foram convertidos para o formato *string*, sendo composto pelo prefixo da cidade - representado pelas 3 primeiras letras do nome da cidade - e o identificador original, separados por underscore "_". Como exemplo desta mudança, o ônibus com identificador "BC941" da cidade de Curitiba teve seu identificador alterado para "CUR_BC941".

4.3. Enriquecimento

A partir da padronização dos dados geoespaciais dos ônibus e de informações das linhas extraídas do site Moovit, foi possível realizar o enriquecimento do *dataset* inserindo dados com informações relevantes acerca das linhas operadas pelos ônibus, os quais são informados a seguir: valor da tarifa; agência que opera a linha e a URL de referência para a linha na plataforma Moovit.

Ademais, também foram coletados manualmente os códigos das cidades utilizados pelo IBGE e incluídos no campo "*city_code*", com o objetivo de realizar a diferenciação entre cidades homônimas, além de facilitar a integração com bases de dados do IBGE e outras fontes de dados.

4.4. Dicionário de dados

O dicionário de dados do *dataset*, exibido na Tabela 2, foi elaborado a partir dos esquemas fornecidos pelas APIs, incluindo os dados descritos na Seção 4.3. Para auxiliar comparações com *datasets* internacionais e a utilização por usuários não falantes da língua portuguesa, a nomenclatura dos campos presentes no BRBus foi redigida em inglês.

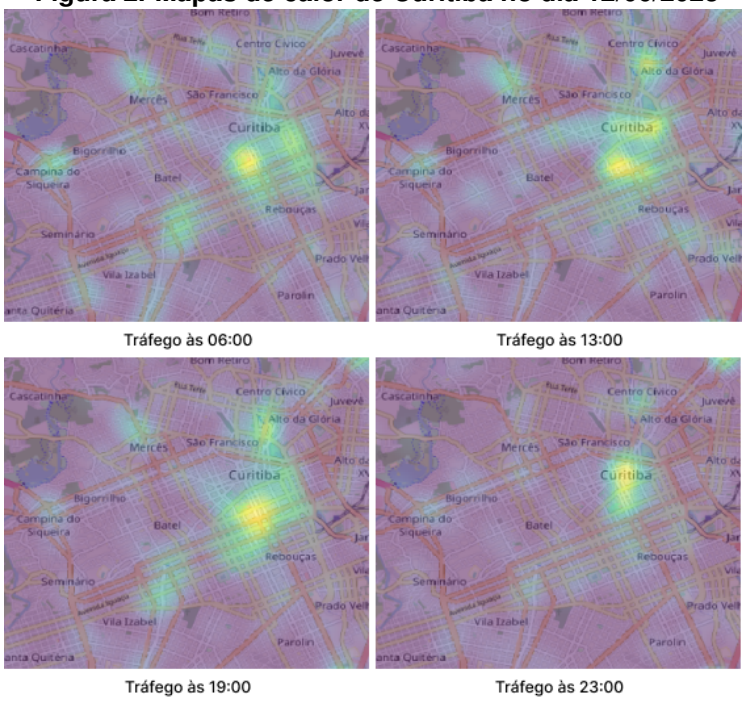
Tabela 2. Dicionário de dados do BRBus, onde os campos em vermelho referem-se a campos adicionados pelos autores

Campo	Descrição	Tipo
bus_id	Identificador do ônibus	String
city	Cidade de operação do ônibus	String
city_code	Identificador da cidade de operação do ônibus na base de dados do IBGE	Integer
agency	Nome da operadora do ônibus	String
line_code	Código da linha do ônibus	String
line_url	URL para a página da operadora da linha (Moovit)	String
line_fare	Tarifa da linha	Float
is_adapted	Verificador de acessibilidade no ônibus	Boolean
is_eletric	Indicador se ônibus é elétrico ou não	Boolean
latitude	Latitude do ônibus no momento da requisição	Float
longitude	Longitude do ônibus no momento da requisição	Float
bus_speed	Velocidade do veículo em Km/h	Float
bus_direction	Sentido da linha que o ônibus está operando	1 = Ida 2 = Volta
updated_at	Data/Hora de obtenção da informação	String (ISO 8601)

5. Aplicações do BRBus

A partir do *dataset* construído neste trabalho, é possível extrair informações relevantes para estudos acerca da mobilidade urbana das cidades utilizadas. Análises como fluxo de

Figura 2. Mapas de calor de Curitiba no dia 12/06/2023



veículos, frota ativa durante o dia, regiões atendidas, tempo de deslocamento, entre outras podem ser úteis para stakeholders como tomadores de decisão de cidades inteligentes, profissionais de TI, entre outros. A seguir, iremos apresentar algumas aplicações baseadas nos dados disponibilizados pelo BRBus.

- **Mapa de calor do tráfego** - Os mapas de calor identificam o volume de dados em uma certa região. Considerando os dados fornecidos pelo BRBus referentes à cidade de Curitiba no dia 12/06/2023, foram gerados mapas de calor da localização dos ônibus, os quais podem ser visualizados na Figura 2. Pode-se observar que a frota de ônibus possui maior concentração na região central da cidade.
- **Visualização do uso da frota durante a semana** - Um tipo de análise que pode ser feita acerca da mobilidade urbana é a obtenção dos veículos em movimento durante um período em relação a frota total disponível, explorando características temporais e geoespaciais dos dados para montar essa análise quantitativa. A Figura 3 contém um gráfico radial com a representação dos picos de utilização da frota de Curitiba a cada hora, onde cada anel representa um dia e cada seção representa as horas do dia. O anel mais interno representa “segunda-feira” e o mais externo “domingo”, considerando a semana entre 12 e 18 de junho de 2023.
- **Comparação entre as velocidades médias dos ônibus** - A velocidade dos ônibus representa uma informação relevante para a análise de tráfego. BRBus disponibiliza as velocidades de cada ônibus, obtidas no momento da requisição, referentes às cidades do Rio de Janeiro e Brasília. Assim, foram calculadas as velocidades médias em diferentes horários do dia 14/06/2023 para ambas as cidades, conforme exibido no gráfico da Figura 4.

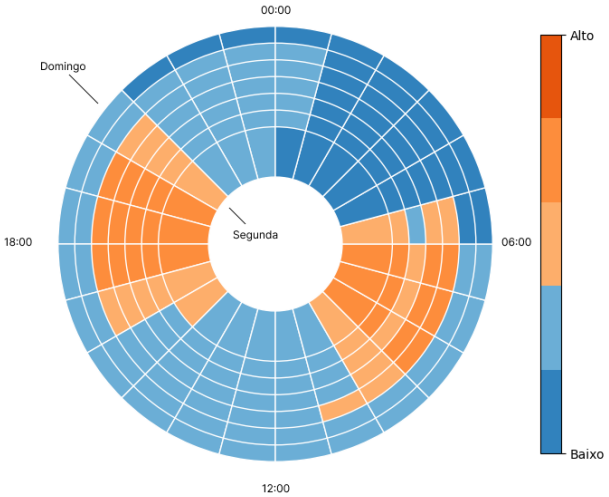


Figura 3. Análise da utilização da frota de Curitiba entre os dias 12/06/2023 e 18/06/2023

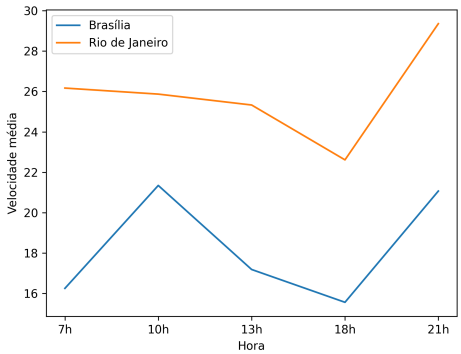


Figura 4. Comparação entre as velocidades médias das cidades do Rio de Janeiro e Brasília no dia 14/06/2023

6. Considerações Finais

Este artigo abordou a construção do *dataset* BRBus: um *dataset* com dados geoespaciais do tráfego de ônibus das cidades de Curitiba, Rio de Janeiro, São Paulo e de Brasília. O *dataset* encontra-se disponível no GitHub, no endereço *https://github.com/brbus-project/dataset* contendo arquivos CSV nomeados com o prefixo da cidade seguido pelo *timestamp* da coleta. Além dos dados, encontra-se no repositório o dicionário de dados utilizado no processo de padronização.

Durante a construção do BRBus foram identificadas algumas limitações nos dados fornecidos pelas cidades, como a carência de informações referentes à condição de acessibilidade dos ônibus (ausentes nas cidades do Rio de Janeiro e Brasília), dados da ocupação de passageiros (não fornecidos por nenhuma cidade) e velocidades dos ônibus (ausentes nas cidades de São Paulo e Curitiba).

Em trabalhos futuros, pretende-se expandir o BRBus com informações de outras cidades, realizar a construção de dashboards para rápidas visualizações dos dados, adição de novos campos relevantes para a análise de dados de trafégo, por exemplo, enriquecendo o dataset com dados do logradouro baseado na localização do ônibus. Também pretende-se permitir a análise em tempo real, por meio do processamento de streaming de dados com ferramentas como Apache Kafka e Apache SparkStreaming, fornecendo informações atualizadas para os tomadores de decisão.

7. Agradecimentos

Este trabalho foi realizado com apoio financeiro da FAPEAL e da UFAL, por meio do Programa Institucional de Bolsas de Iniciação Científica. Também agradecemos a colaboração dos membros do grupo de pesquisa, Ana Wagner e Thiago Emídio, pelas contribuições nas pesquisas que resultaram neste trabalho.

Referências

Alablani, I. and Alenazi, M. (2020). EDTD-SC: An IoT sensor deployment strategy for smart cities. *sensors*, 20(24):7191.

Arbex, R. O. and da Cunha, C. B. (2016). Avaliação das mudanças nas velocidades das linhas de ônibus da cidade de são paulo após a implantação de faixas exclusivas através da análise de dados de gps. *Transportes*, 24(4):21–31.

Braz, T., Maciel, M., Mestre, D. G., Andrade, N., Pires, C. E., Queiroz, A. R., and Santos, V. B. (2018). Estimating inefficiency in bus trip choices from a user perspective with schedule, positioning, and ticketing data. *IEEE Transactions on Intelligent Transportation Systems*, 19(11):3630–3641.

Cruz Junior, J. I. S. d. (2020). Métodos de análise de dados no transporte público urbano. Monografia (Trabalho de Conclusão de Curso) - Ciência da Computação - Universidade Federal de Campina Grande, Campina Grande, 2020.

CSC (2022). Ranking connected smart cities 2022. Disponível em: `https://ranking.connectedsmartcities.com.br/`. [Acessado 20-Jun-2023].

da Cruz, S. M. S., Andrade, L. S., and Sampaio, J. O. (2016). Explorando dados abertos governamentais sobre a mobilidade urbana na cidade do rio de janeiro. In *Anais do XLIII Seminário Integrado de Software e Hardware*, pages 1645–1655. SBC.

de Transporte Público, A. N. (2018). Sistema de informações da mobilidade urbana da associação nacional de transportes público-simob/antp.

Jose, R. and Mitra, S. (2018). Identifying and classifying highway bottlenecks based on spatial and temporal variation of speed. *Journal of Transportation Engineering, Part A: Systems*, 144(12):04018075.

Larsen, G. H., Yoshioka, L. R., and Marte, C. L. (2020). Bus travel times prediction based on real-time traffic data forecast using artificial neural networks. In *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pages 1–6.

Queiroz, A. R. M., Santos, V. B., Nascimento, D. C., and Pires, C. E. S. (2019). Conformity analysis of GTFS routes and bus trajectories. In *Anais do XXXIV Simpósio Brasileiro de Banco de Dados*, pages 199–204. SBC.

Statista (2016). IoT devices installed base worldwide 2015-2025 — Statista — statista.com. `https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/`. [Acessado 02-Jan-2023].

TAO, W. (2013). Interdisciplinary urban GIS for smart cities: advancements and opportunities. *Geo-spatial Information Science*, 16(1):25–34.