QualiBus: implementando métricas de qualidade para dados de transporte coletivo

Rafael Luciano L. Silva[1], **Diêgo de A. Correia**[1], **Letícia A. Mendes**[2], **Ruan T. Melo**[1], **Fábio J. Coutinho**[1]

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

{rlls, dac, rtm, fabio}@ic.ufal.br

[2]Centro de Engenharias e Ciências Agrárias – Universidade Federal de Alagoas (UFAL)

leticia.mendes@ceca.ufal.br

Abstract. *The analysis of public transportation data is essential for planning actions aimed at improving the public transport system in large cities, reducing operational costs, and enhancing urban mobility and user satisfaction. However, analyzing data related to the movement of thousands of vehicles integrated with IoT devices presents challenges regarding the quality and reliability of the information obtained, due to the presence of inconsistencies, inaccuracies, duplicates and gaps. This work proposes a framework for implementing data quality metrics for bus transport, which was validated through a case study using bus movement data from four Brazilian cities.*

Resumo. *A análise de dados de transporte coletivo é fundamental para o planejamento de ações que visem aprimorar o sistema de transporte público em grandes cidades, reduzindo custos de operação e melhorando a mobilidade urbana e satisfação do usuário. Entretanto, analisar dados referentes à movimentação de milhares de veículos integrados com dispositivos IoT impõe desafios relacionados à qualidade e confiabilidade das informações obtidas tendo em vista a presença de inconsistências, imprecisões, duplicidades e lacunas. Este trabalho propõe um framework para a implementação de métricas de qualidade de dados de circulação de ônibus, o qual foi validado a partir de um estudo de caso com dados da movimentação de ônibus de quatro cidades brasileiras.*

1. Introdução

No Brasil, a mobilidade urbana enfrenta importantes desafios visto que grande parte das áreas densamente povoadas carecem de um sistema de transporte público eficiente. Alguns dos problemas conhecidos em grandes cidades brasileiras são a ausência de investimento de transporte de massa e o crescimento do número de acidentes com vítimas, dos congestionamentos e da emissão de poluentes veiculares [de Carvalho 2016]. Na perspectiva de cidades inteligentes, esses desafios podem ser enfrentados a partir da análise de dados dos veículos que integram o sistema de transporte público a fim de obter indicadores que permitam aos gestores planejar ações e promover intervenções no funcionamento do sistema.

Considerando o transporte coletivo realizado por ônibus, a análise de dados de deslocamento dos veículos permite a implementação de ações que visam melhorar aspectos como o monitoramento dos ônibus, a diminuição do tempo de espera dos passageiros, o aumento de cobertura da rede, a redução do tempo de deslocamento, entre outros [Pedrosa 2019] [de Melo et al. 2021]. A maior parte desses dados são provenientes de dispositivos integrados aos veículos e compartilhados através de APIs de tráfego disponibilizadas por órgãos públicos,

provendo informações como localização GPS, velocidade dos veículos e direção do deslocamento. Todavia, é fundamental assegurar a qualidade dos dados para garantir a confiabilidade dos resultados da análise. No entanto, a presença de dados incompletos, medições imprecisas e outras inconsistências podem comprometer significativamente os processos de tomada de decisão. Para superar tal desafio torna-se necessária a implementação de mecanismos para avaliação de métricas de qualidade de dados espaciais, tais como precisão posicional, completude e consistência lógica.

Na literatura, os trabalhos de análise de dados de transporte público evidenciam a necessidade de verificar a qualidade de dados espaciais. [Martins et al. 2022] identificaram problemas nos dados da movimentação de ônibus da cidade de Curitiba, tais como a baixa acurácia no posicionamento de pontos de ônibus, falhas nos dados de localização dos veículos, lacunas em medições e redundâncias. Similarmente, [Yai 2016] apontaram inconsistências nos registros de velocidade e na frequência de atualização de shapes de ônibus da cidade de São Paulo.

A norma ISO 19157-1 (2023) estabelece os princípios utilizados para descrição dos elementos de qualidade de dados espaciais, os quais dividem-se em completude, consistência lógica, precisão posicional, qualidade temporal e qualidade temática. Este trabalho apresenta o QualiBus – um *framework* desenvolvido para avaliar a qualidade de dados de deslocamento de ônibus. Para validar a solução implementada foi realizado um experimento que analisou a qualidade de dados da movimentação de ônibus de quatro grandes cidades brasileiras. Assim, o artigo encontra-se organizado da seguinte maneira: A Seção 2 discute os trabalhos encontrados na literatura com propostas similares; A Seção 3 descreve os conceitos, padrões e métricas que serviram de base para este trabalho; A Seção 4 apresenta o *framework*; A Seção 5 descreve experimentos realizados para avaliar o impacto da qualidade de dados em diferentes aplicações; A Seção 6 apresenta a síntese dos resultados, destacando suas implicações práticas, bem como recomendações e propostas para trabalhos futuros.

2. Trabalhos Relacionados

A coleta e o processamento de dados de tráfego em tempo real fornecem dados essenciais para as operações do sistema de transporte público. No entanto, a utilidade dessas informações depende diretamente da qualidade dos dados coletados e da aplicação de padrões que permitam sua análise, avaliação e enriquecimento. Alguns trabalhos na literatura propõem soluções para lidar com a verificação da qualidade de dados de transporte público, os quais são brevemente discutidos a seguir.

Segundo [dos Santos et al. 2016], a acurácia posicional dos dados geoespaciais puntiformes pode ser otimizada. Para isso, utilizaram um cenário simulado, com base nos referenciais da norma brasileira ET-ADGV (2010), empregando os métodos da "estatística do vizinho mais próximo de alta ordem", que calcula a distância entre os pontos definidos, e da Função K de Ripley, que avalia a variância total entre os pontos definidos. Esses métodos permitem, assim, a análise da tendência dos dados e a avaliação da sua precisão.

[Diniz Junior et al. 2017] revelaram importantes inconsistências no sistema de Curitiba, propondo soluções para correções e melhorias, com destaque a recomendação da padronização dos dados - garantindo maior confiabilidade e uniformidade nas informações. Outrossim, foram implementadas distintas formas de visualização de dados geoespaciais, como grafos e mapas, oportunizando análises mais aprofundada voltadas ao gerenciamento de tráfego e itinerário.

Enquanto o estudo de [Yunus et al. 2017] propõe um *framework* focado em seis dimensões essenciais – precisão, completude, inconsistência, relevância, acessibilidade e atuali-

dade – aplicado no sistema SIKAP da SPAD, o presente trabalho adota a perspectiva da ISO 19157-1:2023 para identificar inconsistências nos dados das APIs públicas de quatro cidades brasileiras.

Apesar dos avanços promovidos pela comunidade científica no estudo dos dados geoespaciais, ainda existem lacunas relevantes quanto à realização de análises mais abrangentes e integradas sobre a qualidade das informações disponibilizadas por APIs públicas do transporte coletivo.

3. Qualidade de dados de transporte público

A compreensão sobre o significado de qualidade de dados pode ser resumida em garantir que o dado é adequado ao uso (*"fit for use"*) ou que atende aos requisitos de seus usuários [Wang and Strong 1996] [Redman 2001]. Os requisitos de qualidade dos dados podem ser definidos por padrões, legislação, regulamentos, políticas, partes interessadas ou seu uso pretendido [Fürber 2015]. No contexto de dados espaciais, o *framework* de qualidade deve utilizar métricas e padrões que avaliem tanto a corretude de informações georreferenciadas (coordenadas e topologias) quanto a consistência dos atributos vinculados às localizações, os quais são necessários para aferir o nível em que os dados estão livres de falhas e se adéquam ao seu propósito.

Em se tratando de dados de transporte coletivo, a qualidade dos dados permite que as análises sejam conduzidas com base em informações com maior nível de consistência, favorecendo o processo de tomada de decisão dos gestores públicos. Por exemplo, em cidades inteligentes, quando os registros de localização dos veículos atendem aos requisitos esperados, sistemas voltados para planejamento de rotas, previsão de demandas e monitoramento em tempo real alcançam maior precisão, melhorando a qualidade de vida dos usuários do sistema de transporte público.

Neste trabalho, os requisitos, critérios e métricas de qualidade de dados espaciais orientam-se pela norma ISO 19157-1:2023 (*Geographic Information - Data Quality*) e pelo padrão *General Transit Feed Specification* (GTFS), ambos reconhecidos internacionalmente e descritos nas subseções a seguir.

3.1. ISO 19157-1:2023 - Qualidade de Dados Geoespaciais

A ISO 19157-1:2023 representa o principal normativo que estabelece requisitos gerais para a descrição e avaliação da qualidade de dados geoespaciais. Ela provê uma estrutura sistemática de componentes e métricas de qualidade, definindo procedimentos padronizados para quantificar atributos de dados espaciais, sendo subdividida em:

- **Completude:** Indica se todas as informações esperadas e necessárias estão presentes no conjunto de dados, sendo particionada em **comissão** (excesso de dados) e **omissão** (ausência de dados).
- **Consistência lógica:** Refere-se ao grau em que os dados aderem às regras estruturais, semânticas e relacionais que os governam. Sendo dividida em: **(i) consistência conceitual**: aderência dos dados com as restrições conceituais e semânticas do esquema; **(ii) consistência de domínio**: validação de que os valores dos atributos respeitam os domínios pré-definidos, como tipo e intervalo; **(iii) consistência de formato**: adequação dos valores ao formato definido para o conjunto de dados; **(iv) consistência topológica**: manutenção das relações espaciais e estruturais entre as feições, garantindo que a representação geográfica obedeça a topologia.

- **Qualidade temporal:** Trata da qualidade dos atributos temporais e das relações temporais das feições presentes nos dados geográficos. Divide-se em: **(i) validade temporal**: validez dos dados em relação ao tempo; **(ii) consistência temporal**: assegura a ordem lógica e a integridade das sequências de eventos; **(iii) acurácia temporal**: exatidão das referências temporais e a incerteza associada.
- **Acurácia posicional:** Descreve a proximidade entre a localização de uma feição geográfica conforme reportada no conjunto de dados e a sua verdadeira localização no mundo real.
- **Acurácia temática:** Mede a correção dos valores quantitativos registrados e a consistência das classes atribuídas às feições, comparando-os com valores de referência (verdadeiros). Inclui subelementos como corretude de classificação, acurácia de atributos não quantitativos e acurácia de atributos quantitativos.

3.2. Padrão GTFS

O padrão GTFS é uma especificação de formato aberto e universal para representação de dados de transporte público, projetado para facilitar o acesso e a interoperabilidade de informações como rotas, horários, paradas e tarifas. Dividido em GTFS *Schedule* (dados estáticos) e GTFS *Realtime* (dados dinâmicos), o padrão permite que agências de transporte publiquem informações estruturadas em arquivos de texto simples (CSV) e atualizações em tempo real via *Protocol Buffers*. Enquanto o GTFS *Schedule* especifica a representação de dados da operação regular (como itinerários e geolocalização de pontos), o GTFS *Realtime* é uma extensão do padrão GTFS utilizada por agências de transporte público para representar alertas de serviço, posições de veículos e ajustes de viagens, facilitando a troca de informações atualizadas em cenários dinâmicos, desde que alimentadas por fontes confiáveis e sistemas de monitoramento adequados.

4. O *framework* QualiBus

A escassez de soluções para avaliação da qualidade de dados de transporte público foi identificada na revisão da literatura conforme discutido na Seção 2. Nesse sentido, foi realizado um levantamento acerca dos principais requisitos envolvidos no contexto de dados geoespaciais referentes ao sistema de transporte coletivo. A partir disso, foi desenvolvido um *framework*, denominado QualiBus, para avaliar as métricas de qualidade em dados do fluxo de ônibus considerando a ISO 19157-1 e o formato GTFS.

O *framework* tem como propósito principal facilitar a integração de práticas de avaliação da qualidade de dados no desenvolvimento de soluções, promovendo maior confiabilidade, transparência e eficácia na utilização de dados geoespaciais para planejamento, operação e monitoramento do transporte público. Ao fornecer um processo direcionado às demandas do sistema de transporte público utilizando ônibus, QualiBus disponibiliza um conjunto de métricas de qualidade baseadas em normas e padrões, visando reduzir inconsistências e imprecisões na análise de dados da movimentação de ônibus.

4.1. As etapas de verificação das métricas de qualidade

Na Figura 1 são apresentados os módulos do *workflow* do QualiBus, cada um correspondendo a uma etapa fundamental da avaliação da qualidade de dados em conformidade com as métricas da ISO 19157-1 apresentadas anteriormente na Seção 3.1. A execução das etapas segue a ordem do fluxo dos dados a partir de seu carregamento.

Inicialmente, a fim de encontrar possíveis problemas estruturais na base de dados avaliada, é realizada uma análise de completude. Os campos são examinados considerando dois

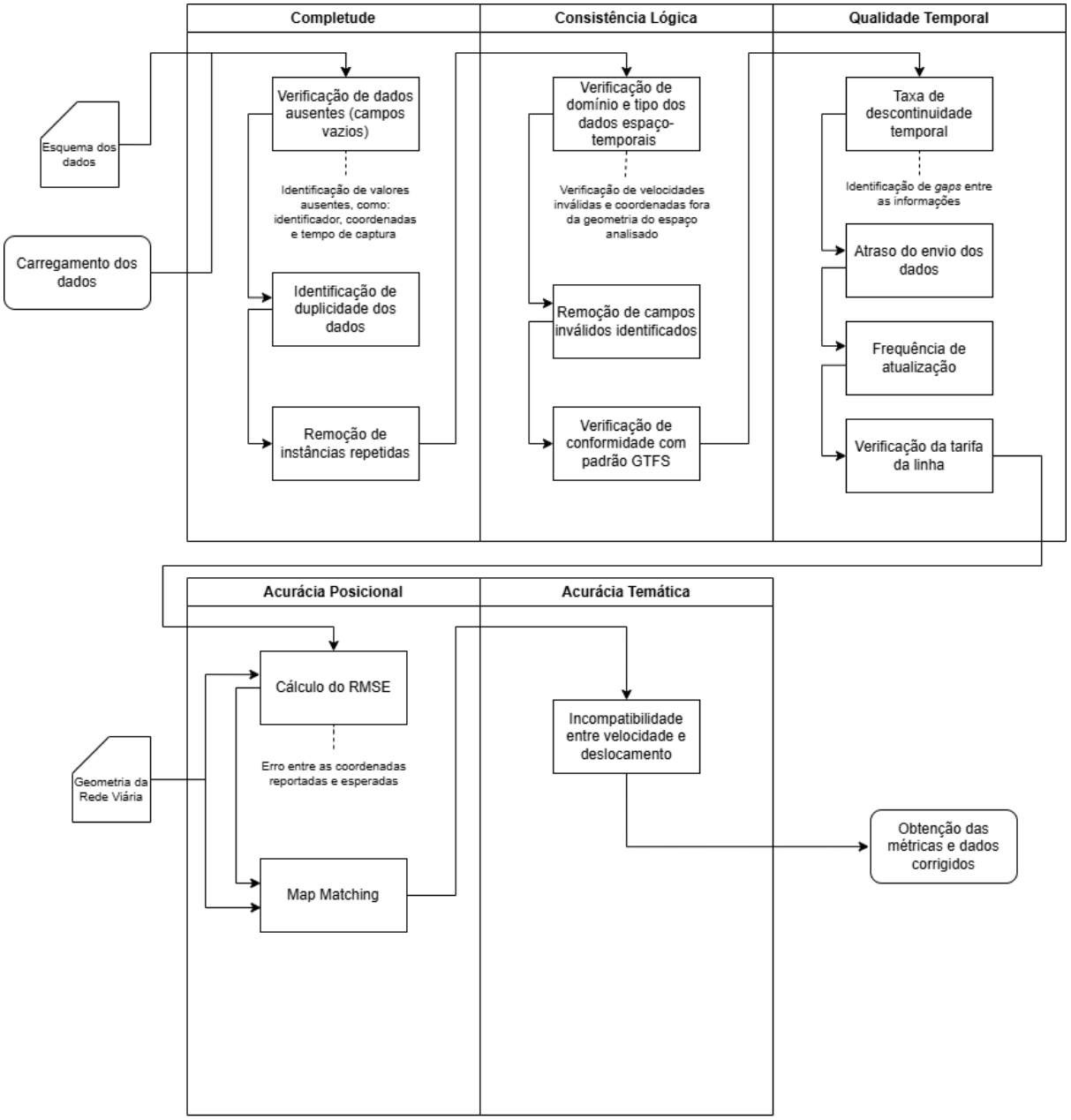


Figura 1. O *workflow* do QualiBus

principais critérios: omissão e comissão. O critério de omissão é verificado na base de dados avaliada a partir da detecção, contagem e registro da métrica de ausência de dados (campos vazios). Dessa forma, para cada campo da base de dados, executa-se uma varredura ao longo das instâncias coletadas, no intuito de identificar a presença de campos vazios, proceder a contagem e produzir a métrica.

O critério de comissão é verificado a partir da identificação de dados de localização duplicados, evidenciando um possível problema com o funcionamento do dispositivos do veículo ou no sistema de compartilhamento da agência de transporte. Para detecção da duplicidade é realizado um agrupamento dos campos de id e *timestamp* de uma determinada instância e analisado se esse grupo se repete em outras instâncias da base. Após serem detectadas e contabilizadas, as instâncias duplicadas são removidas, mantendo-se apenas uma e eliminando as repetidas.

Após a análise da completude, inicia-se a verificação da consistência lógica dos dados avaliados. Nesta etapa, verifica-se a conformidade com padrões estabelecidos, como o GTFS, analisando os domínios e tipos dos dados espaço-temporais (*e.g.* velocidades dos veículos e limites territoriais abrangidos pelo serviço). A checagem do domínio inicia por verificar se os valores de cada campo estão de acordo com o esquema. Do mesmo modo, o tipo passa por uma checagem, onde é comparado se o tipo presente no campo é o mesmo esperado no esquema do conjunto. Diante da detecção destes dados inválidos, caso o campo seja crítico (id, coordenadas e *timestamp*), remove-se completamente a captura, caso não, remove-se apenas o valor do campo. Por fim, é verificado se os padrões de formatação nos campos são os mesmos estabelecidos previamente, como *timestamps* fora do formato de data e hora da ISO 8601-1:2019.

A verificação da qualidade temporal inicia a partir da avaliação da taxa de descontinuidade temporal, que se caracteriza como uma detecção de uma possível interrupção nas medições de um mesmo ônibus. Essa interrupção é definida por um limiar para o intervalo entre *timestamps*, o qual pode ser estabelecido pelo usuário do QualiBus ou obtido na documentação da API da agência de transporte público (fonte dos dados avaliados). Caso não seja configurado um limiar, o QualiBus utiliza a regra estatística dos três sigmas, para proceder o seu cálculo conforme descrito nos passos a seguir: (i) Agrupa os dados por id do ônibus; (ii) Para cada ônibus, ordena os dados por *timestamp*; (iii) Calcula os intervalos entre os *timestamps* de cada ônibus; (iv) Obtém a média aritmética e desvio padrão dos intervalos; (v) Calcula o limar utilizando a regra dos três sigmas(média + 3 desvios padrão).

Durante a verificação da validade temporal, o *framework* considera as seguintes métricas: atraso no envio das medições, frequência de atualização e a verificação da tarifa da linha. O atraso no envio dos dados pode ser obtido a partir da análise do tempo médio entre a captura dos dados e sua disponibilização via API da agência de transporte. A frequência de atualização é definida a partir do cálculo da média aritmética dos intervalos de tempo entre os *timestamps* para um mesmo id de ônibus, fazendo uma média global utilizando as médias individuais em seguida. Para sua aferição são implementados os seguintes passos: (i) Os dados são agrupados por id e ordenados por seus *timestamps*; (ii) Para cada ônibus são calculados: os intervalos entre seus *timestamps* em segundos e a média aritmética desses intervalos, armazenando-se essa informação de média por id; (iii) Com as informações armazenadas se calcula a média individual para a obtenção da média global; (iv) É disponibilizado o arquivo com a média por ônibus e outro com a média global, obtendo assim o diagnóstico dessa métrica globalmente e localmente, visando identificar possíveis problemas em ids específicos.

Além disso, a verificação da validade temporal dos valores da tarifa para cada linha de ônibus pode ser realizada a partir de uma base de referência, sendo obtida através da comparação dos *timestamps* associados a cada valor de tarifa da referência com a base de dados analisada, checando a validez desse valor para o período de tempo observado.

Para a obtenção da acurácia posicional verifica-se o quão precisas são as localizações reportadas pelos dispositivos GPS em relação às suas posições verdadeiras. Esta verificação é feita através do índice RMSE (Root Mean Square Error), que calcula a diferença quadrática média entre as coordenadas emitidas e as esperadas. Posteriormente, a técnica de Map-Matching pode ser aplicada para a correção do desvio posicional dos dados. Esta técnica baseia-se em fazer a correspondência dos pontos reportados com uma rede de mapas digitais, associando essas coordenadas às vias ou trajetos mais prováveis. Para fazer a implementação do Map-Matching, recomenda-se a utilização de ferramentas *open-source* com grande capacidade de processamento, como Fast Map Matching [Yang and Gidofalvi 2018] ou Barefoot

[Mattheis et al. 2014].

A verificação da acurácia temática é realizada a partir da implementação de um método para a checagem de velocidades zeradas, conferindo se os valores do campo referente à velocidade, na sequência ordenada de capturas de cada ônibus em um determinado dia, permanecem zerados à medida em que a posição geográfica e o *timestamp* mudam, o que poderia indicar desvio em relação às condições operacionais esperadas ou falhas na coleta/emissão dos dados.

A Tabela 1 apresenta uma síntese das métricas da ISO 19157-1 selecionadas para aferir a qualidade dos dados de transporte coletivo:

ISO 19157-1	**Métricas**	**Descrição**	**Tipo/Unidade**
Completude	Taxa de ausência dos dados (omissão)	Percentual de registros com campos vazios	Percentual (%)
	Detecção de duplicatas (comissão)	Identificação de registros redundantes (mesmo ID e *timestamp*)	Percentual (%)
Consistência lógica	Verificação de domínio	Identificação de valores fora de intervalos esperados	Contagem / Percentual (%)
	Adesão ao padrão de formato	Checa os tipos e formatos (Ex.: *timestamp* no formato da ISO 8601)	Qualitativo (válido/inválido)
Qualidade temporal	Frequência de atualização	Intervalo médio entre capturas consecutivas para o mesmo veículo	Segundos (s)
	Atraso no envio das medições	Tempo entre a captura dos dados e a disponibilização via API	Segundos (s)
	Taxa de descontinuidade temporal	Interrupção nas medições de um ônibus	Segundos (s) / Percentual (%)
	Verificação de tarifa	Conformidade com valores de tarifa esperados	Qualitativo (válido/inválido)
Acurácia posicional	RMSE (Root Mean Square Error)	Erro médio entre coordenadas GPS capturadas e esperadas	Metros (m)
Acurácia temática	Acurácia de atributos quantitativos: frequência de velocidades zeradas	Percentual de registros com velocidades igual a zero na ausência do dado	Percentual (%)

Tabela 1. Métricas de qualidade baseadas na ISO 19157-1:2023

4.2. Implementação

O QualiBus é implementado em Python, utilizando a interface de Python para Apache Spark, *Pyspark*[1]. Essa abordagem foi escolhida para garantir, desde sua concepção, a capacidade para escalar horizontalmente para grandes volumes de dados. A implementação em Python também favorece e facilita a integração com a biblioteca Pandas[2], que é amplamente utilizada por analistas de dados. Pela natureza adaptativa do Apache Spark, o *framework* pode ser executado localmente de forma direta, sem necessidade de alteração de código, assim como de forma distribuída, ficando a cargo do usuário fazer apenas a configuração do ambiente do

[1]https://www.databricks.com/br/glossary/pyspark
[2]https://pandas.pydata.org/

cluster. O código-fonte do QualiBus se encontra disponível no Github, a partir do endereço https://github.com/QualiBus/Algoritmos.

5. Experimento

Para validar o funcionamento do QualiBus foi realizado um experimento em duas etapas. Inicialmente, utiliza-se o QualiBus para verificar a qualidade de um *dataset* de monitoramento geoespacial de ônibus de 4 cidades brasileiras[3]. Em seguida, desenvolve-se uma aplicação para demonstrar o impacto da qualidade de dados na estimativa do tempo de espera de passageiros em paradas de ônibus. As seções a seguir descrevem em detalhes as fases do experimento.

5.1. Métricas de Qualidade usando o QualiBus

Nesta primeira parte do experimento, um *dataset* é carregado no QualiBus a fim de obter as métricas de qualidade. Apesar do *framework* permitir o processamento distribuído, este experimento foi realizado em uma única máquina (*single-node*). O *dataset* utilizado é formado por dados coletados a partir de agências de transporte público de cidades brasileiras, fornecendo informações espaço-temporais que refletem a realidade operacional dos ônibus no Brasil e permitem avaliar métricas de qualidade. Os dados foram produzidos por sistemas de rastreamento GPS instalados em ônibus das seguintes cidades: Brasília, Curitiba, Rio de Janeiro e São Paulo.

O *dataset* inclui atributos como localização do veículo (latitude e longitude), *timestamp* (tempo de captura), velocidade (disponível apenas nos dados de Brasília e Rio de Janeiro), identificador do ônibus (único por veículo) e informações da linha. Os dados foram coletados em um período de 9 dias e totalizam 84GB. As etapas de coleta e desenvolvimento do *dataset* estão descritas em [Melo et al. 2023].

Considerando a métrica Completude, QualiBus verificou para cada campo do dataset a existência de valores vazios. Devido à limitação de espaço, apresenta-se aqui apenas a taxa de ausência de dados do campo velocidade. A verificação desta métrica pelo QualiBus aplica-se às linhas de ônibus de forma individualizada, podendo revelar o quantitativo de linhas que possuem cobertura de dispositivos que registram a velocidade dos veículos. Assim, QualiBus apresenta o resultado desta métrica agrupando-o em 3 categorias conforme descrito a seguir:

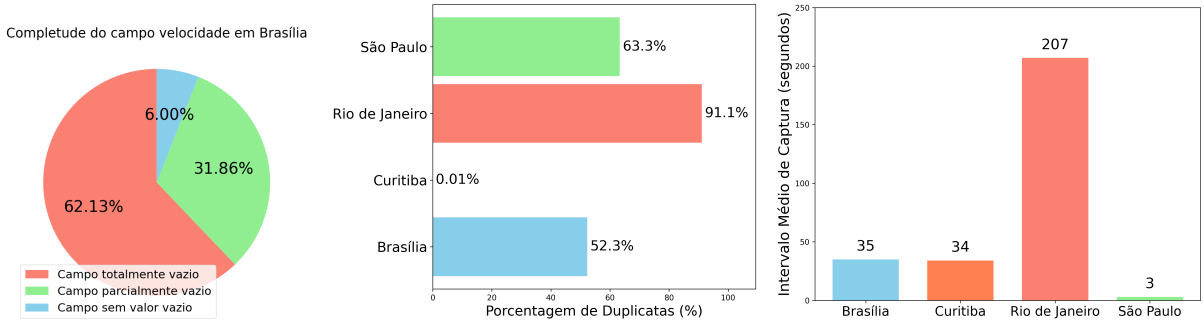
- Campo totalmente vazio: linhas em que o campo velocidade aparece vazio para todas as capturas.
- Campo parcialmente vazio: linhas em que parte das capturas apresenta o campo velocidade vazio.
- Campo sem valor vazio: linhas em que o campo velocidade aparece não-vazio para todas as capturas.

Para a cidade do Rio de Janeiro não foi detectada, para o campo velocidade, a existência de valores vazios considerando todas as linhas de ônibus. A Figura 2a apresenta um gráfico representado o índice de completude do campo velocidade para a cidade de Brasília. A métrica expressa que 62.13% das linhas de ônibus da cidade possuem o campo velocidade vazio em todas as leituras, representando uma possível ausência de medição em mais da metade da rede de cobertura. QualiBus revelou também que 6.0% das linhas possuem a velocidade parcialmente vazias, enquanto 31.86% das linhas não exibiram valores nulos em nenhuma leitura.

A métrica de completude foi verificada para a detecção de dados duplicados (medições repetidas) e os resultados são exibidos na Figura 2b, que destaca a variabilidade na proporção

[3]https://github.com/brbus-project/dataset

de capturas repetidas encontradas nas cidades. A cidade de Curitiba apresentou um baixo índice dados repetidos (0.01%) enquanto as demais cidades mais da metade de suas medições eram repetidas, sendo 91.1% no Rio de Janeiro, 63.3% de São Paulo e 52.3% de Brasília. A repetição de dados de captura na cidade do Rio de Janeiro apresentou um patamar elevado, fato também mencionado em [Rojas and Lanzilloti 2019].



(a) (b) (c)

Figura 2. (a) Proporção de velocidades nulas nas linhas de Brasília. (b) Proporção de duplicatas nas 4 cidades. (c) Frequência de atualização nas 4 cidades.

Na métrica de consistência lógica, apenas o método de verificação de domínio foi escolhido. A investigação se deu por determinar a quantidade de velocidades que estavam acima de 110km/h (valor escolhido por representar o limite de velocidade máxima de um veículo segundo o Código de Transito Brasileiro[4]). O *dataset* não apresentou velocidades fora do domínio definido, indicando uma forte consistência lógica dentro dos parâmetros definidos.

Considerando a métrica de qualidade temporal, QualiBus retornou a frequência de atualização das capturas dos ônibus. Os resultados dessa métrica constatam um valor significativamente maior para a cidade do Rio de Janeiro, com um intervalo médio de 207 segundos, destoando de Brasília, com 35 segundos, seguida por Curitiba com 34 segundos e São Paulo, que indica 3 segundos de intervalo médio entre captura. Esta variação pode ser melhor observada na Figura 2c.

Para a métrica de acurácia temática, QualiBus retornou a porcentagem de velocidades zeradas, apresentando valores discrepantes para as cidades de Brasília e Rio de Janeiro. A tabela 2 evidencia que no Rio de Janeiro, 61.2% dos registros de velocidade são zerados, contrastando com os 30.2% encontrado em Brasília. Além disso, 83.2% das linhas de ônibus do Rio de Janeiro continham mais da metade de velocidades zeradas, contra 7.8% das linhas de Brasília.

Tabela 2. Proporção de velocidades zeradas para Brasília e Rio de Janeiro.

Cidade	Velocidades Zeradas	Linhas com >50% de Capturas Zeradas
Brasília	30.2%	7.8%
Rio de Janeiro	61.2%	83.2%

Levando em consideração os resultados extraídos das métricas de completude e acurácia temática, pode-se deduzir que a ausência de valores nulos para o campo velocidade na cidade do Rio de Janeiro pode ser explicado por um possível comportamento de sua API, atribuindo valor zero para a ausência de medição, ocasionando um problema de domínio.

[4]https://www.ctbdigital.com.br/artigo/art61

5.2. Impacto da Qualidade de Dados em Análises

Para demonstrar o impacto da qualidade de dados em análises reais, foi desenvolvida uma aplicação para estimar o tempo de espera de passageiros em pontos de parada de ônibus, sendo utilizado os dados do BRBus e as malhas viárias das cidades de Brasília[5] e Rio de Janeiro[6]. Estas cidades foram escolhidas para o experimento por serem as únicas que disponibilizam os valores de velocidade.

Em um processamento inicial, os dados do *dataset* foram separados por linhas de ônibus, onde cada arquivo representa os trajetos dos ônibus de uma determinada linha.

Posteriormente, para cada linha, foi feito um *matching* dos pontos de parada em relação às vias, utilizando a abordagem *snap-to-nearest*. Dessa forma, são analisados apenas os pontos de paradas que fazem parte das rotas de cada linha.

Inicialmente, para cada ponto de parada, foram selecionados os dois ônibus distintos, pertencentes à linha analisada, que estivessem mais próximos do ponto. Os intervalos do tempo de captura desses ônibus não deve ser superior a 60 minutos a fim de evitar a inclusão de ônibus registrados em intervalos de tempo muito distantes (possível descontinuidade temporal), o que poderia distorcer as estimativas.

A distância entre os dois ônibus mais próximos é utilizada para calcular o tempo em que o segundo levaria para alcançar a posição do primeiro. Dessa forma, este intervalo representa o tempo estimado entre a passagem dos dois veículos pela parada.

Na Figura 3a podem-se observar os ônibus de diferentes linhas, retratados pelos pontos coloridos, onde cada cor corresponde a uma linha diferente, e a distância traçada entre os dois ônibus de uma mesma linha mais próximos da parada, que será usada para fazer a estimativa final.

Algumas limitações foram encontradas durante a análise dos ônibus mais próximos da parada. Os problemas de acurácia do GPS impossibilitaram que a distância fosse calculada a partir das vias percorridas. A possibilidade de utilizar o *map-matching* não foi considerada visto que adicionaria uma camada de complexidade na aplicação, impedindo a verificação da influência da acurácia posicional, perdendo-se um importante ponto de observação e impacto. Assim, foi utilizada a distância em linha reta entre os ônibus.

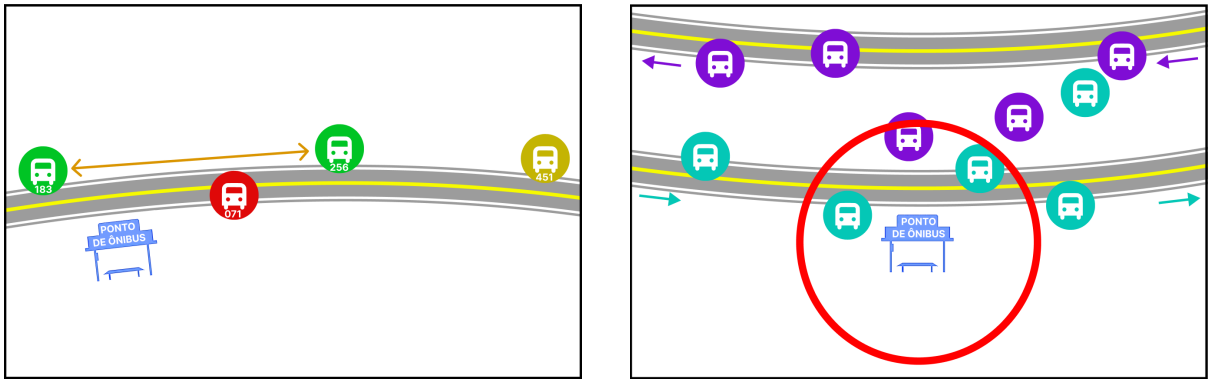
Outro desafio foi identificar corretamente os ônibus mais próximos que trafegavam na mesma via que a parada. Em cenários com vias de mão dupla, a seleção dos ônibus mais próximos poderia ser inconsistente. Por exemplo, um ônibus considerado próximo à parada de uma via poderia estar circulando na via adjacente e caso fosse utilizado, alteraria o resultado da estimativa.

Para superar esse obstáculo, foi utilizado um algoritmo de *K-Nearest Neighbors* (KNN). Esse algoritmo analisa as direções dos ônibus localizados dentro de um raio de 6 metros da parada, e determina a direção mais frequente entre eles como a direção da parada.

A Figura 3b representa o cenário descrito anteriormente, onde as cores ciano e roxo refletem ônibus no sentido de ida e volta e o círculo vermelho expressa o raio de busca.

[5]https://geoserver.semob.df.gov.br/geoserver/web/

[6]https://www.data.rio/datasets/



(a) (b)

Figura 3. (a) Estimativa do tempo de espera entre 2 ônibus de uma mesma linha mais próximos da parada. (b) KNN aplicado ao ponto de parada, considerando um raio de 6 metros, para determinar a direção da via em que ele se encontra.

5.2.1. Resultados e Discussões

Os resultados da aplicação desenvolvida para estimar o tempo de espera de passageiros em pontos de parada de ônibus foram obtidos para as cidades de Brasília e Rio de Janeiro. Para o Rio de Janeiro, o tempo médio de espera calculado foi de 69 minutos. Em Brasília, o tempo médio de espera foi significativamente menor, registrando cerca de 23 minutos.

Os dados estimados para Brasília condizem com o tempo de espera médio disponibilizado pela plataforma Moovit[7], evidenciando um indicativo de coerência com outras fontes, diferente dos valores estimados para o Rio de Janeiro, que contrastam com os 21 minutos divulgados pela plataforma. Para compreender as razões dessa diferença e avaliar as estimativas, foram utilizados os seguintes critérios descritos no QualiBus: a frequência de atualização dos dados e a acurácia temática na variável de velocidade.

Frequência de atualização dos dados: Os resultados apresentados na Seção 5 permitem inferir que a diferença da frequência de atualização impacta diretamente a precisão das estimativas, pois à medida em que a frequência de atualizações diminui, geram-se maiores lacunas temporais nas trajetórias e distâncias espaciais maiores entre os ônibus. Assim, esta relação entre a baixa frequência de atualização e a distância média entre ônibus pode agravar a imprecisão das estimativas. Por exemplo, nos dados do Rio de Janeiro, a distância média é de 595,27 metros, correspondendo a mais que o dobro dos 257,89 metros em Brasília. Tal fato confirma que a distância influencia os cálculos da aplicação, já que é utilizada para aferir o tempo de movimentação entre os dois ônibus mais próximos.

Acurácia temática na variável de velocidade: Velocidades zeradas podem refletir situações reais, como paradas ou congestionamentos. Entretanto, uma proporção elevada, como apontada na Seção 5, pode indicar falhas na emissão de dados, especialmente se o ônibus estiver em movimento no momento da captura. Na aplicação desenvolvida, a velocidade média diária do segundo ônibus mais próximo da parada é utilizada para calcular o tempo de sua chegada à posição do primeiro ônibus. Quando esta velocidade média é registrada próxima a zero, o cálculo pode resultar em tempos de espera inconsistentes. Assim, a alta incidência de velocidades zeradas no Rio de Janeiro pode ser um fator contribuinte para a estimativa de tempos de espera mais longos.

[7]https://moovitapp.com/insights/en/Moovit_Insights_Public_Transit_Index-countries

A disparidade encontrada nos resultados das estimativas do tempo de espera para Brasília e Rio de Janeiro evidencia como a qualidade dos dados pode impactar nas análises. Tal fato reforça a necessidade de melhorias na emissão de dados e a observação das normas de padronização tal como a ISO 19157-1 a fim de aprimorar a confiabilidade dos dados.

6. Considerações Finais

Este artigo apresentou o desenvolvimento do QualiBus, um *framework* voltado à implementação de métricas de qualidade para dados de deslocamento de ônibus. Foi realizado uma revisão da literatura que evidenciou a necessidade de soluções para tratar a qualidade de dados no contexto de sistema de transporte público. Neste contexto, foi verificado que a ISO 19157-1 e o padrão GTFS são importantes normativos para orientar o processo de verificação da qualidade de dados de transporte público, o quais serviram de base para o QualiBus. O *framework* foi desenvolvido para atender a 10 métricas de qualidade, as quais foram validadas a partir de um experimento realizado com dados reais quatro cidades brasileiras. Os resultados das métricas evidenciam o impacto da qualidade de dados em análises do sistema de transporte público.

Como trabalhos futuros, pretende-se implementar técnicas para a verificação de dados coletados em tempo real (*e.g.* checagem dos trajetos dos ônibus), além da integração com outras aplicações para a obtenção de dados complementares tais como condições de tráfego e eventos urbanos.

Este trabalho foi realizado com apoio financeiro da UFAL, por meio do Programa Institucional de Bolsas de Iniciação Científica.

Referências

de Carvalho, C. H. R. (2016). Desafios da mobilidade urbana no brasil. Technical report, Texto para discussão.

de Melo, L. A., Gonzalez, L. F. G., and Borin, J. F. (2021). Identificando e recuperando anomalias em sistemas de veículos com rotas. In *Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP)*, pages 81–90. SBC.

Diniz Junior, P. C. et al. (2017). Serviços telemáticos em uma rede de transporte público baseados em veículos conectados e dados abertos. Master's thesis, Universidade Tecnológica Federal do Paraná.

dos Santos, A. d. P., Rodrigues, D. D., Santos, N. T., and Gripp, J. (2016). Avaliação da acurácia posicional em dados espaciais utilizando técnicas de estatística espacial: Proposta de método e exemplo utilizando a norma brasileira. *Boletim de Ciências Geodésicas*, 22(4):630–650.

Fürber, C. (2015). Semantic technologies. In *Data quality management with semantic technologies*, pages 56–68. Springer.

Martins, T. S. et al. (2022). Map matching: Uma análise de dados streaming de trajetórias de gps no transporte público. *Anais Estendidos do XVIII Simpósio Brasileiro de Sistemas de Informação (SBSI 2022)*.

Mattheis, S., Al-Zahid, K. K., Engelmann, B., Hildisch, A., Holder, S., Lazarevych, O., Mohr, D., Sedlmeier, F., and Zinck, R. (2014). Putting the car on the map: a scalable map matching system for the open source community. In *Informatik 2014*, pages 2109–2119. Gesellschaft für Informatik eV.

Melo, R. T., Vasconcelos, F. F., Silva, R. L. L., Santos, P. V., Ramos, V. T., and Coutinho, F. J. (2023). BRBus - construindo um dataset para monitoramento geoespacial dos ônibus de cidades brasileiras. In *Dataset Showcase Workshop (DSW)*, pages 25–35. SBC.

Pedrosa, L. (2019). Técnicas para detecção de anomalias em padrões de séries espaço-temporais: Uma revisão sistemática de literatura. *ETIS-Journal of Engineering, Technology, Innovation and Sustainability*, 1(1):41–53.

Redman, T. C. (2001). *Data quality: the field guide*. Digital press.

Rojas, A. and Lanzilloti, R. (2019). Análise da velocidade operacional dos ônibus durante os jogos olímpicos e paraolímpicos Rio 2016 baseados em dados de GPS. *Cadernos do IME - Série Informática*, 40:43–58.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33.

Yai, A. K. (2016). Análise e visualização de dados do transporte público de ônibus da cidade de São Paulo.

Yang, C. and Gidofalvi, G. (2018). Fast map matching, an algorithm integrating hidden markov model with precomputation. *International Journal of Geographical Information Science*, 32(3):547–570.

Yunus, F. M., Magalingam, P., Maarop, N., Samy, G. N., Hooi-Ten Wong, D., Shanmugam, B., and Perumal, S. (2017). Proposed data quality evaluation method for a transportation agency. *Open International Journal of Informatics (OIJI)*, 5(2):52–63.