Quando chega? Análise de previsibilidade de tempos de espera em transporte público urbano utilizando Modelos Mistos

Brendon L. Pasquim[1], **Keiko V.O. Fonseca**[1], **Luiz Ledo M. Melo Jr.**[1]

[1]Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial (CPGEI)
Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de Setembro, 3165 – 80230-901 – Curitiba – PR – Brazil

brendonpasquim@alunos.utfpr.edu.br,{keiko,luizledo}@utfpr.edu.br

Abstract. *This paper presents an exploratory analysis of the use of city open data to find scientific evidences for urban mobility public policy updating or creation. The results leaded to possible applications on heuristics and strategies to processing and analyzing data from public transportation as well related variables. The contributions of the study include the design and implementation of mixed models, the integration of methods and tools specifically designed to handle public transportation data; a discussion about how the model results can be used to assess the quality of service of the bus line operation; and recommendations on the usage of city open data to foster public policies to increase the adoption of public transportation.*

Resumo. *Este artigo apresenta uma análise exploratória do uso de dados abertos da cidade para encontrar evidências científicas para atualização ou criação de políticas públicas de mobilidade urbana. Os resultados levaram a possíveis aplicações em heurísticas e estratégias para processamento e análise de dados de transporte público, bem como variáveis relacionadas. As contribuições do estudo incluem o design e a implementação de modelos mistos, a integração de métodos e ferramentas especificamente projetados para lidar com dados de transporte público; uma discussão sobre como os resultados do modelo podem ser usados para avaliar a qualidade do serviço da operação da linha de ônibus; e recomendações sobre o uso de dados abertos da cidade para fomentar políticas públicas para aumento da adoção do transporte público.*

1. Introdução

Os efeitos das mudanças climáticas têm demandado estudos para o transporte público de massa [Wimbadi et al. 2021] [Gallivan et al. 2011]. A adoção do usuário pelo transporte público depende não só da sua condição financeira, mas também da sua percepção sobre a relação custo-benefício. Se apresentar custo e benefícios melhores ou equivalentes comparada com o uso do veículo individual aumenta a chance do usuário decidir pelo uso do sistema de transporte público urbano (STPU) [Habib et al. 2011]. Em termos de benefícios, em geral se considera a qualidade da prestação do serviço, por exemplo, medida pelo tempo de deslocamento, acessibilidade, comodidade, conforto e usabilidade, pontualidade (aqui entendida como aderência à tabela horária fornecida previamente), previsibilidade (aqui entendida como o estudo da redução da variabilidade prevista com a aproximação temporal de um evento previsto)[Büchel and Corman 2022], entre outros [Habib et al. 2011].

A partir de um estudo exploratório, busca-se contribuir com (1) um melhor entendimento sobre a operação do STPU a partir de padrões nos dados abertos disponíveis de uma cidade, (2) o uso de modelos para a análise da pontualidade e previsibilidade da prestação de serviços do STPU; (3) e predições específicas de serviços inovadores do STPU bem como novas hipóteses a serem investigadas em estudos confirmatórios. Um estudo de caso com dados abertos de Curitiba, em especial, de ônibus urbanos do seu STPU é apresentado e os resultados discutidos à luz da qualidade de dados, modelos e estudos anteriores. Em específico, considera-se a tabela horária do STPU a partir de informações geotemporais públicas e avalia-se a aderência dessa tabela a partir de dados históricos. Investiga-se ainda outros possíveis dados públicos que possam ser relacionados ao sistema de transporte público e sua operação, possíveis fatores de impacto na aderência à tabela horária, bem como estratégias de melhor previsão da chegada de ônibus em seus pontos de parada.

Este documento está assim organizado: a partir desta introdução, a próxima seção apresenta os conceitos chave de STPU, de modelos mistos e de ferramentas de manipulação de dados aplicados neste trabalho, e uma revisão da literatura. A Seção 3 descreve o método utilizado e detalha o estudo de caso; e a Seção 4 traz os resultados obtidos, uma abordagem implementada para sua validação e uma discussão sobre os mesmos. A Seção 5 apresenta as considerações finais e alguns estudos futuros.

2. Conceitos básicos para o entendimento das contribuições

2.1. Dados Públicos Abertos

Adota-se aqui o conceito de dados abertos como aqueles disponíveis de forma livre, podendo ser utilizados ou republicados sem restrições de direitos autorais ou patentes [Braunschweig et al. 2012]. Algumas cidades oferecem dados abertos do Transporte Público tais como Linhas e Pontos de ônibus, itinerários e a geolocalização dos ônibus para consulta e consumo. Neste trabalho utiliza-se dados abertos fornecidos pela URBS (Urbanização de Curitiba) e IPPUC (Instituto de Pesquisa e Planejamento Urbano de Curitiba)[1] descritos na seção 3.2.

2.2. Sistemas de transporte público

Transporte público pode ser definido como qualquer forma de transporte de passageiros ou cargas disponível para aluguel e recompensa. Geralmente se refere ao transporte terrestre de passageiros através de ônibus, trens e suas variantes [Preston 2009]. Em geral, um sistema de transporte público tem sua qualidade associada aos serviços prestados, ao número de usuários atendidos, seu custo, além de seus impactos sociais e ambientais.

O presente estudo limita-se ao transporte público de passageiros em ônibus urbanos, sem perda de generalidade. Define-se aqui itinerário ou percurso de um veículo associado a uma rota e horários pré-determinados como uma sequência ordenada de pontos que o veículo deve obedecer para embarque e desembarque de passageiros. Passageiros são aqueles que utilizam os veículos de transporte público para se deslocar. Pontos de

[1] `https://dadosabertos.c3sl.ufpr.br/curitiba/TransporteColetivo` (acesso em 01/04/2025)

Ônibus ou Pontos de Parada de Ônibus são locais dedicados para o embarque e desembarque de passageiros. Por fim, um ônibus é um veículo automotor de grande porte dedicado ao transporte em massa de passageiros.

A avaliação de desempenho do serviço de transporte público, em geral, considera o comportamento temporal da oferta do serviço ao usuário. Embora o desempenho de STPU possa ser associado a vários parâmetros, entre eles, ocupação do ônibus, aderência à tabela horária, conforto interno, frequência do serviço, etc; nosso estudo foca na avaliação da previsibilidade a partir dos dados abertos disponíveis.

2.3. Modelos Mistos

Em muitas áreas a investigação científica busca estabelecer um modelo geral para uma população, onde dentre as várias características em comum dessa população, seleciona-se uma variável de interesse Y e um vetor de variáveis explicativas independentes $X = [X_1, \ldots, X_n]$ com objetivo de predizer e explicar Y. Um possível candidato de modelo geral é o da regressão múltipla, definida por:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

onde $\epsilon \sim N(0, \sigma^2)$ e N uma distribuição normal de média zero e variância σ^2. βs são os coeficientes angulares em relação à variável explicativa independente e o ϵ os fatores residuais (valores não previstos pelo modelo), mais os possíveis erros de medição

Note que se faz um suposição de homocedasticidade, ou seja, assume-se que em $Y|x \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2)$, a variância é constante para qualquer valor x e foi fixada para toda população. Esta é uma suposição assumida ao longo da população descartando a possibilidade da formação de grupos na população que podem ter diferença na variabilidade [Matloff 2017].

Uma forma de levar em conta a diferença entre grupos na população é com o uso de modelos mistos, uma metodologia que incorpora a regressão múltipla como efeito fixo da população e também permite que os coeficientes e interceptos β_0 variem em função dos grupos [Faraway 2006].

Em aplicações onde ao longo do tempo ocorrem medidas repetidas de indivíduos ou grupos na população sob análise, é possível estabelecer uma análise de dados longitudinais [Fitzmaurice et al. 2012]. Um caso particular de modelos mistos é aqui definido como:

$$Y_{it} = \beta_0 + b_{0i} + \beta_1 t_{1it} + b_{1i} t_{1it} + \cdots + \beta_p t_{pit} + b_{pi} t_{pit} + \epsilon_{it}$$

onde $i = 1, ..., G$, representam o grupo; $t = 1, \ldots, T$, representam o tempo; e o total de observações $N = G \times T$, t_{pit} é valor da variável explicativa *p* do grupo *i* no tempo *t*. b_{0i}, b_{pit} são os efeitos aleatórios para o intercepto β_0 e a inclinação específicos do indivíduo e para o erro $\epsilon_{it} \sim N(0, \sigma^2)$. No modelo acima, os termos b_{0i} e b_{pi} representam efeitos aleatórios específicos do grupo i. O termo b_{0i} afeta o intercepto β_0, ou seja, permite que cada grupo tenha um valor médio diferente. Já os termos b_{pi} afetam a inclinação de cada variável explicativa t_{pit}, permitindo que os coeficientes variem entre os grupos.

Nosso estudo apresenta duas propostas de modelagem do erro (ϵ_{it}) do grupo *i* no tempo *t* baseada na existência de correlação temporal entre os dados.

$$Corr(\epsilon_{it}, \epsilon_{ik}) = 0, \forall \quad t \neq k \tag{1}$$

A equação 1 modela o erro quando não existe correlação temporal entre os dados que são avaliados através dos resíduos. Caso contrário, isso é, se houver uma estrutura temporal dos resíduos, então:

$$Corr(\epsilon_{it}, \epsilon_{ik}) = \rho\sigma^2, \forall \quad t \neq k \tag{2}$$

Identificando a estrutura de correlação como na equação 2 pode-se ajustar modelos Box e Jenkins [Box et al. 2015].

No presente estudo, utiliza-se Modelos Mistos para modelar o comportamento do STPU, no qual o efeito fixo aliado ao efeito aleatório modelam as diversas variáveis que afetam o tempo de espera em pontos de parada de ônibus (por exemplo, chuva, congestionamento, embarque e desembarque de passageiros em pontos de parada de grande fluxo, acidentes de trânsito, etc). Neste contexto, β_0 representa a média geral de tempo de espera (em segundos), isto é, de um mesmo ônibus para todos os pontos de ônibus em um dado itinerário. Ou seja, é um valor constante dentro do itinerário.

Já b_{0i} representa o fator aleatório específico de cada ponto, sendo responsável por ponderar o tempo de espera de cada ponto de ônibus em relação a média (sendo um número positivo ou negativo). Por isso, é um valor variável por ponto de ônibus do itinerário. Um valor positivo significa que o tempo de espera no ponto de parada é superior a média geral de espera do itinerário. Da mesma forma, um valor negativo representa um tempo de espera inferior a média geral do itinerário.

2.4. Correlograma e seleção de modelos

Correlograma é a representação gráfica da correlação entre múltiplas variáveis (correlação cruzada), ou até mesmo, da mesma variável ao longo do tempo (autocorrelação). No contexto da análise de autocorrelação o correlograma é também conhecido como "gráfico de correlação serial". Dentre as diversas aplicações de correlogramas estão o aprimoramento de modelos e resíduos, análise de dependência em séries temporais e escolha de modelos de previsão. No presente trabalho, correlogramas foram utilizados para definir parâmetros autorregressivos para resíduos de modelos de predição. Para isso, calculou-se a correlação entre observações defasadas por j períodos de tempo para se obter os coeficientes de autocorrelação (equação 3):

$$r_j = \frac{\sum_{t=1}^{n-j}(t_{i'k} - \bar{t}_{i'k})(t_{i'k+j} - \bar{t}_{i'k})}{\sum_{t=1}^{n}(t_{i'k} - \bar{t}_{i'k})^2} \tag{3}$$

Os coeficientes de autocorrelação são dados por $r_j = \frac{c_j}{c_0}$, onde c_j é a covariância entre observações defasadas em j períodos de tempo. Um gráfico com os j primeiros coeficientes de autocorrelação como função de j é chamado de correlograma. Modelos autorregressivos tem correlogramas que caem exponencialmente [Box et al. 2015].

A seleção do modelo mais adequado (ou seja, aquele com melhor ajuste aos dados reais) se fez através do "Erro Quadrático Médio" (EQM). EQM é uma métrica usada para avaliar a qualidade de previsões em modelos estatísticos, medindo a diferença média quadrática entre os valores previstos $\hat{t_{i'k}}$ e observados $t_{i'k}$. EQM é definido na equação 4:

$$EQM = \frac{1}{n}\sum_{t=1}^{n}(t_{i'k} - \hat{t_{i'k}})^2 \tag{4}$$

2.5. Revisão da Literatura

Um indicador associado à qualidade do serviço de transporte público é a aderência à tabela horária (caracterizada por instantes de tempo de chegada e partida dos veículos do transporte público em um ponto específico de seu itinerário), oferecendo ao usuário previsibilidade para embarcar, desembarcar, fazer conexões e assim alcançar seu destino em um intervalo de tempo limitado [Büchel and Corman 2022].

A previsibilidade advinda da predição do tempo de chegada é afetada pela qualidade e abrangência dos dados reais utilizados. Erros, incertezas e dados repetidos no conjunto de dados afetam os resultados das previsões, exigindo a aplicação proativa de estratégias de tratamento, como em [Martins et al. 2022] ou [Hashiguchi et al. 2020].

Modelos para predição de tempos de chegada já foram propostos anteriormente [Suwardo et al. 2010], [Dong et al. 2013], [Li et al. 2017], [Kumar et al. 2025]. Entretanto, utilizam um conjunto de dados pequeno, são dependentes de dados em tempo-real, custosos em termos de recursos computacionais ou negligenciam os desafios inerentes aos dados de transporte. [Hashiguchi et al. 2020] investiga a cidade de Curitiba especificamente, contribuindo ao indicar diferenças no padrão do horário de pico, como considerado pelo operador do STPU, mas não oferece informações de tempo de espera em paradas.

[Curzel et al. 2019] propõe uma metodologia baseada em *link streams* para modelar e analisar temporalmente a operação de STPU com dados reais de Curitiba. A análise considera um terminal de ônibus e suas conexões entre linhas mas não as possíveis causas de acúmulo de atrasos na operação das linhas.

O presente trabalho busca superar limitações dos artigos estudados ao analisar em detalhes a estrutura dos dados do STPU, bem como tratá-los de forma estruturada através de [Peixoto et al. 2020], identificar itinerários com [Borges et al. 2023] e implementar um modelo misto baseado em dados offline históricos, aplicando-o a um estudo de caso na cidade de Curitiba.

3. Metodologia

A análise do comportamento temporal de ônibus de transporte público utilizando dados abertos e modelos mistos se baseou em avaliar a aderência dos modelos propostos aos dados reais de operação de um STPU. Pela disponibilidade de dados e ferramentas de processamento específicas para o conjunto de dados existente, o sistema integrado de transporte público de Curitiba foi escolhido.

O método consistiu em:

1. Estudar um conjunto limitado de dados de transporte público para um período de tempo específico e linha de ônibus arbitrária;
2. Identificar itinerários utilizando o algoritmo proposto em [Borges et al. 2023];
3. Computar, através de modelos mistos, a média geral de tempo de espera para todos os pontos de ônibus e ponderar o tempo de espera de cada ponto em relação à média para uma linha e itinerário de ônibus arbitrário;
4. Validar os resultados do modelo de previsão em relação aos dados reais, em uma data arbitrário, avaliando sua aderência.

O item 1 objetivou coletar, limpar e normalizar um conjunto reduzido de dados de transporte público de Curitiba, permitindo assim analisar sua estrutura. O item 2 expandiu o item 1, resolvendo problemas relacionados a lacunas e inconsistências nos dados através de uma estratégia robusta para a identificação de itinerários.

O item 3 buscou analisar em detalhes o comportamento do itinerário através de modelos mistos identificando quais pontos de ônibus possuem tempos de espera reduzidos ou elevados e o quanto estes se distanciam da média geral de espera. Por fim, o item 4 objetivou aplicar o modelo de previsão na mesma linha modelada, mas em uma data arbitrária distinta, avaliando se este produziu resultados satisfatórios para os tempos de espera em pontos de parada de ônibus.

3.1. Formalização

Entende-se aqui por Itinerário um conjunto S de termos ordenados s_i onde cada s refere-se a geolocalização de um ponto de ônibus (localização pré-especificada para embarque e desembarque de passageiros no ônibus) e i o número de sequência do ponto no percurso, formando assim uma sequência ordenada de pontos de ônibus definida em (5):

$$S = (s_1, s_2, s_3, ..., s_n) \tag{5}$$

Em 5, $i, n \epsilon \mathbb{N}_{>0}$ sendo que n representa o último ponto de ônibus do percurso. A localização geográfica dos pontos de ônibus, bem como a ordem de parada em cada ponto na rota definida, são determinados pelo operador do STPU.

Ainda, dado o Itinerário S e os pontos de ônibus s_i e s_j, se define T como o Tempo de Espera do Itinerário, expresso na equação 6, como uma sequência de tempos de espera $t_{i \to j}$ entre cada par de pontos de ônibus, onde $t \epsilon \mathbb{R}_{>0}$ e $i, j \epsilon \mathbb{N}^*$, sendo $i < j$, $j \leq n$.

$$T = (t_{1 \to 2}, t_{2 \to 3}, t_{3 \to 4}, ..., t_{(n-1) \to n}) \tag{6}$$

A partir de 6 calcula-se $Tempo$, o tempo total necessário para um ônibus completar um Itinerário. Esse valor é expresso na equação 7:

$$Tempo = \sum_{\substack{i=1 \\ j=i+1}}^{n-1} t_{i \to j} \tag{7}$$

A Figura 1 ilustra os conceitos de *Itinerário* e *Tempo de Espera do Itinerário*:

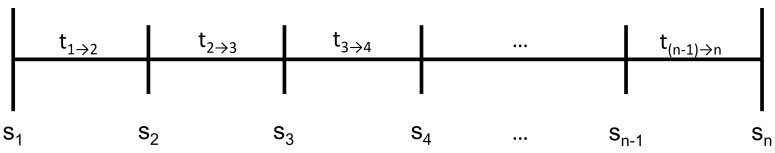

Figura 1. Conceitos de Itinerário S e Tempo de Espera de Itinerário T

3.2. Dados do sistema de Transporte Público

Como exposto na seção 2.1, a cidade de Curitiba publica dados brutos de transporte público como dados abertos. Tais dados são reunidos e armazenados pelo C3SL[2] . Os dados disponibilizados são:

- **Veículos** - As coordenadas geográficas dos veículos nas linhas de transporte;
- TrechosItinerarios - Os trechos dos itinerários das linhas de transporte;
- TabelaVeiculo - O número da tabela horária que o veículo estava executando;
- TabelaLinha - A tabela horária das linhas de transporte;
- ShapeLinha - As coordenadas por onde as linhas de transporte passam;
- **PontosLinha** - Os pontos de ônibus que pertencem as linhas de transporte;
- Pois - Pontos de referência da Cidade de Curitiba;
- **Linhas** - As linhas da Rede Integrada do Transporte Coletivo de Curitiba.

Os itens em negrito são aqueles necessários para se identificar itinerários de ônibus. "Veículos", "PontosLinha" e "Linhas" contém dados relacionados a posição geográfica dos ônibus no itinerário, a posição geográfica e endereço dos pontos de parada e, o nome e identificador numérico *ID* da linha, respectivamente. Estes dados, quando ainda não processados, são aqui denominados "dados brutos".

Os dados brutos são disponibilizados em D+1, ou seja, um dia após serem coletados pelo STPU. Estes dados estão empacotados e comprimidos em formato *tar.gz*. Após descomprimidos, os dados ficam disponíveis em formato *json*. Nota-se que os dados de geolocalização dos ônibus também são disponibilizados em tempo-real, a cada 2 minutos, através de um *webservice*[3]. Entretanto, se optou por utilizar os dados em D+1 pela sua granularidade, visto que estão mapeados a cada 20 segundos, em média.

O processo de baixar e preparar os dados brutos em um formato adequado para análise foi efetuado em diversos trabalhos anteriores, como em [Hashiguchi et al. 2020] e [Manika 2022], que utilizam o formato JSON. Já o formato parquet, que dispõe os dados em colunas e é otimizado para compressão e recuperação, foi empregado em [Borges et al. 2023] e [Peixoto et al. 2020].

Este trabalho utilizou o formato parquet para o processamento dos dados públicos de Curitiba, obtendo-os através da aplicação "urbs-data-processing"[4], do mesmo autor de [Peixoto et al. 2020]. Com os dados de transporte em formato parquet, se iniciou o processo de identificação de itinerários, descrito na seção 3.3.

3.3. Identificação de itinerários

Identificar itinerários a partir dos dados brutos não é trivial: não há uma relação clara entre a posição geográfica (geolocalização) de um ônibus e do ponto de parada pelo qual ele vai passar, em um dado momento. Além disso, os dados podem apresentar lacunas, não havendo informações disponíveis em certo intervalo de tempo.

[2] `https://www.c3sl.ufpr.br/` (acesso em 22/02/2025)

[3] `https://dadosabertos.c3sl.ufpr.br/curitiba/TransporteColetivo/Documentao_WEB-SERVICE__TRANSPORTE_COLETIVO_DE_CURITIBA.pdf` (acesso em 10/06/2024)

[4] `https://github.com/altierispeixoto/urbs-data-processing.git` (acesso em 20/04/2025)

Esse trabalho utilizou a aplicação "busanalysis" desenvolvida em [Borges et al. 2023] para identificar itinerários. A aplicação executa três passos: 1. Associa a posição geográfica de pontos de parada com ônibus, identificando o horário (*Map Matching*); 2. Ordena os registros a partir do horário (Sequenciamento Temporal); 3. Associa a sequência temporal com um itinerário.

Como entrada, três arquivos de dados em formato parquet são necessários: 1. Veículos; 2. PontosLinha; 3. Linhas (já citados em 3.2). Além disso, é necessário informar a data e o Código ID da linha de ônibus a ser processada.

Como saída, a aplicação produz um arquivo CSV (*Comma-separated Values*) contendo: 1. Código ID da Linha; 2. Código ID do Veículo; 3. Código ID do Ponto de Parada; 4. Código ID do Itinerário ativo; 5. Horário de passagem do ônibus pelo ponto; 6. Número de sequência do ponto dentro do itinerário;

Lacunas nos dados brutos podem ocorrer devido a problemas de comunicação dos dados de geolocalização. Por isso, durante o passo 3 a aplicação implementa interpolações e ajustes de temporização quando necessário, garantindo uma sequência de horários sem falhas. O código fonte pode ser consultado no repositório Git "busanalysis"[5]. Por fim, o arquivo CSV de saída é utilizado como entrada para o script R que implementa o Modelo Misto proposto. Este modelo é detalhado na seção 3.4.

3.4. Predição de Tempo de Espera com Modelos Mistos

O modelo misto proposto tem como objetivo gerar previsões para o tempo de espera de usuários do transporte público em pontos de parada. O Tempo de Espera é definido na equação 6. Este modelo é implementado através de um script R[6].

O script R implementado[7] recebe como entrada o Código do Itinerário, o Código do Ônibus e o arquivo CSV gerado durante o processo descrito na seção 3.3. Como saída, um relatório em formato HTML é gerado, contendo valores de tempo de espera médio e ponderado por ponto, além de representações gráficas dos tempos de espera total por ponto e previsões considerando intervalos de confiança. O pseudocódigo é também disponibilizado no repositório git.

A linha Convencional 863-Água Verde foi escolhida para avaliação. Esta trafega entre os bairros Centro e Água Verde da cidade de Curitiba, escolhida por cruzar a região central. Salienta-se que outras linhas (como 323-V. Autódromo) e datas (26/06/2024, 20/11/2024 e 22/01/2025) foram testadas, sem mudança significativa nos resultados.

A linha 863 tem dois itinerários para dias úteis, 12494 (Sentido Praça Tiradentes/Centro) e 12477 (Sentido Água Verde). Na data de 10/05/2023 (Terça-Feira) dois ônibus estavam em operação, HA310 e HC303 (operando entre 5h da manhã e meia-noite). O veículo HA310 e itinerário 12494 foram escolhidos para análise.

A Figura 2 representa o método usado desde a coleta dos dados brutos até o cálculo das previsões de tempo de espera. As setas cinzas iniciadas com uma linha vertical representam parâmetros arbitrários.

[5] `https://github.com/jcnborges/busanalysis.git` (acesso em 20/04/2025)

[6] `https://www.r-project.org/about.html` (acesso em 12/01/2025)

[7] `https://github.com/Brendonpasquim/mixed-model-bus-wait-time-forecast` (acesso em 20/04/2025)

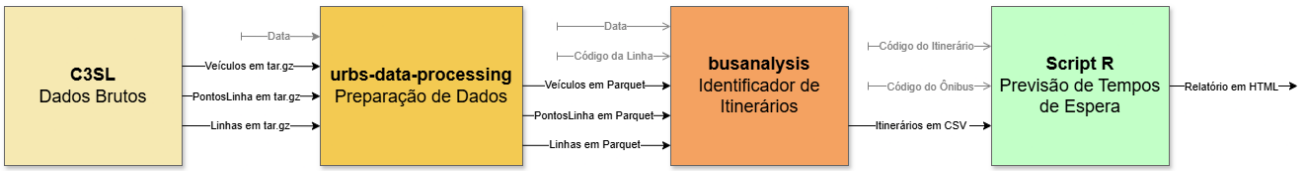

Figura 2. Diagrama de blocos indicando etapas e dados de entrada/saída.

4. Discussão e Resultados

Três modelos de predição foram desenvolvidos de forma incremental, no qual novos componentes foram adicionados para melhor lidar com os dados reais.

4.1. Modelos de previsão

Ao se comparar os três modelos para estimar o tempo de espera do Itinerário (t_{ij}), denomina-se $t_{ij} = t_{i'}$, sendo $i^{'}$ a representação de um efeito aleatório, ou seja, representa o deslocamento de um ponto i para um ponto j no itinerário.

- Modelo 1 (M1): $t_{i'k} = \beta_0 + \epsilon_{t_{i'k}}$, onde $\epsilon_{t_{i'k}} \sim N(0, \sigma^2)$, $i^{'} = 1, \ldots, n$, $k = 1, \ldots, K_{i'}$, sendo n o número de deslocamentos e $K_{i'}$ número de replicações do deslocamento $i^{'}$. Assume-se um mesmo efeito fixo (média geral) para todos os tempos de espera.
- Modelo 2 (M2): $t_{i'k} = \beta_0 + b_{0i'} + \epsilon_{t_{i'k}}$, $\epsilon_{t_{i'k}} \sim N(0, \sigma^2)$, $i^{'} = 1, \ldots, n$, $k = 1, \ldots, K_{i'}$, sendo n o número de deslocamentos e $K_{i'}$ o número de replicações do deslocamento $i^{'}$. Assumimos um mesmo efeito fixo (média geral) e adicionamos mais um efeito aleatório representado por um intercepto por cada tempo de espera relacionado com o respectivo deslocamento.
- Modelo 3 (M3): $t_{i'k} = \beta_0 + b_{0i'} + \epsilon_{t_{i'k}}$, $\epsilon_{t_{i'k}} = \theta_1 \epsilon_{t_{i'(k-1)}} + \cdots + \theta_r \epsilon_{t_{i'(k-r)}} + \Theta_1 \epsilon_{t_{i'(k-s)}} + \cdots + \Theta_p \epsilon_{t_{i'(k-ps)}} + \nu_{t_{i'k}}, \nu_{t_{i'k}} \sim N(0, \sigma_e^2)$, $i^{'} = 1, \ldots, n$, $k = 1, \ldots, K_{i'}$, sendo n o número de deslocamentos e $K_{i'}$ o número de replicações do deslocamento $i^{'}$. Assumimos um mesmo efeito fixo (média geral), adicionamos um efeito aleatório representado por um intercepto para cada tempo de espera relacionado com o respectivo deslocamento e incorporamos uma modelo autorregressivo nos resíduos (estimativa do erro).

  No componente autorregressivo temos o vetor de parâmetros $(\theta_1, \ldots, \theta_r)$ contendo valores entre 0 e 1, que controlam a dependência do resíduo estimado no tempo t em relação aos resíduos estimados no tempo $t-1, \ldots, t-r$.

  O vetor de parâmetros $(\Theta_1, \ldots, \Theta_p)$ avalia se ocorre uma dependência do resíduo no tempo t de maneira sazonal ou cíclica, defasados $t-s, t-ps$, unidades no tempo. Para efeito de notação, caso $(\Theta_1, \ldots, \Theta_p)$ seja não significativo o modelo M3 desconsidera-se a sazonalidade, caso contrário, temos um M3 sazonal.

A Figura 3 apresenta um exemplo de previsão para tempos de espera considerando um dia completo de operação, ou seja, as N vezes que um ônibus executou um itinerário S. Observa-se que utilizando apenas M1 assume-se um único valor de espera para todos, o que não representa a realidade dos dados. Em M2, o intercepto individualizado por cada tempo de espera adicionado ao efeito fixo se ajusta melhor aos dados.

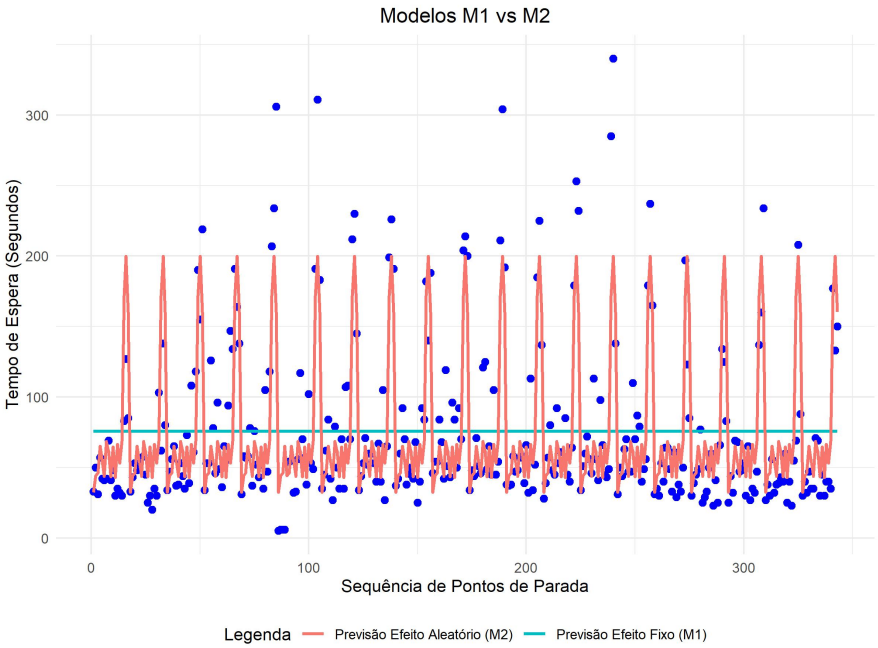

Figura 3. Comparação de previsões dos modelos M1 e M2 considerando todos os N itinerários S de um dia completo de operação do veículo.

4.2. Validação dos modelos

Após o ajuste do modelo é importante a validação através dos resíduos. Para avaliar a qualidade de ajuste do modelo se utilizou um correlograma nos resíduos. Caso o modelo esteja bem ajustado o correlograma não terá nenhuma correlação significativa, ou seja, fora dos intervalos de confiança (correlograma de ruído branco) [Faraway 2006].

Ao analisar-se o modelo M2 se notou nos resíduos que a maior correlação está na primeira defasagem, indicando que os tempos de espera são influenciados pelos tempo de espera imediatamente anterior. Também é possível notar outras correlações significativas acima do intervalo de confiança.

Na Figura 4 (Esquerda) aplicou-se um correlograma no modelo M3. Se nota que boa parte dos valores de correlação se acomodaram dentro do intervalo de confiança. Entretanto, uma correlação significativa e aparentemente sazonal fora do intervalo de confiança persistiu (*Lag 16*). Para modelar essa correlação, se aplicou um modelo autorregressivo também ao componente sazonal. A Figura 4 (Direita) mostra o resíduo do modelo M3 já acrescido do componente sazonal no *Lag 16*. Nota-se que o correlograma apresentou o comportamento esperado, dentro do intervalo de confiança.

Esse componente sazonal no ponto de ônibus de sequência 16 (portando, o *Lag 16*) demonstra uma característica particular em relação aos demais pontos do itinerário 12494. Ao analisá-lo, se percebe que este ponto é uma praça no centro de Curitiba que possui um papel essencial para as conexões com outras linhas e, mesmo não possuindo um terminal de integração espacial, permite integração temporal da tarifa de transporte em determinadas linhas de ônibus.

Esse papel especial justifica o tempo de espera mais elevado, bem como sua recorrência, em comparação com os demais pontos de parada dentro do itinerário, visto que,

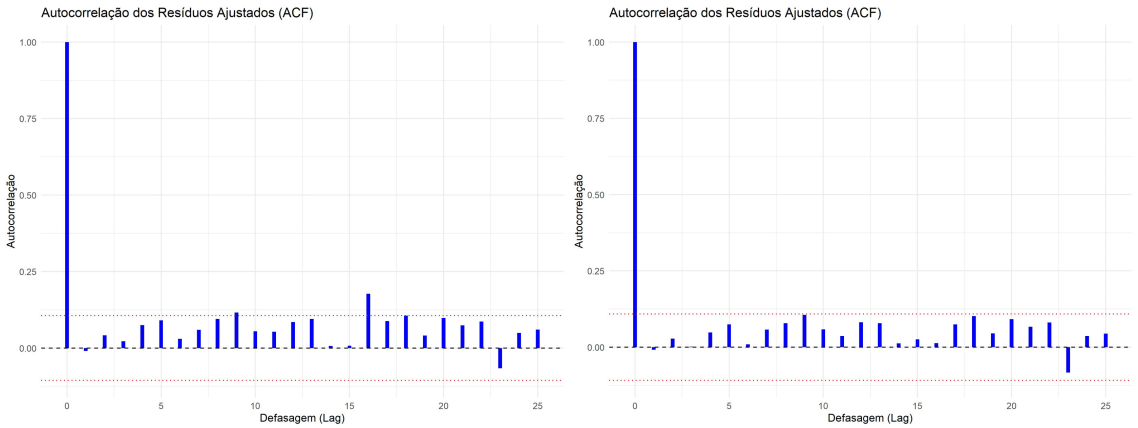

Figura 4. Gráfico de Autocorrelação do modelo M3 (Esquerda) e M3 com sazonalidade (Direita).

um maior número de embarques e desembarques de passageiros é esperado.

4.3. Análise de propriedades dos resíduos

A análise da série residual foi realizada para avaliar a estacionariedade e a presença de autocorrelação significativa. Primeiramente, aplicou-se o **teste de Ljung-Box** para verificar autocorrelações significativas nos resíduos. O teste apresentou um valor de $X^2 = 11.544$, com 10 graus de liberdade ($df = 10$), resultando em um valor de $p = 0.3167$. Como o valor-p é maior que o nível de significância de 0,05, não há evidências suficientes para rejeitar a hipótese nula de ausência de autocorrelação significativa.

Em seguida, foi aplicado o **teste de Dickey-Fuller aumentado (ADF)** para verificar a estacionariedade dos resíduos. O teste apresentou um valor estatístico de -5.9879, com ordem de defasagem $k = 6$, e um valor p de 0,01. Isso indica a rejeição da hipótese nula de não estacionariedade em favor da hipótese alternativa, sugerindo que a série residual é estacionária. Em resumo, os resíduos não apresentam autocorrelação significativa e são estacionários, de acordo com os testes aplicados.

Na figura 5 compara-se M2 e M3 para um itinerário S (de S_1 até S_{18} e $t_{1\to2}$ até $t_{17\to18}$, vide equação 6). Foi possível constatar que os dois últimos deslocamentos, $t_{16\to17}$ e $t_{17\to18}$, possuem uma maior variabilidade. O modelo M3 conseguiu ter mais valores no intervalo de confiança. O modelo M3 adicionado com a sazonalidade não produziu resultados melhores que M3 pela avaliação gráfica.

Na Tabela 1 estão os valores da métrica EQM obtida para os 3 modelos implementados (inclui-se também a variante sazonal de M3). A partir dela se seleciona o modelo "M3 Sazonal" como aquele que apresentou melhor performance.

Os parâmetros para o modelo "M3 Sazonal" são expressos na equação 8. Nela temos $\beta_0 = 75.65$ e $\epsilon_{t_{i'k}} = 0.2842\epsilon_{t_{i'k-1}} + 0.1828\epsilon_{t_{i'k-16}}$. $b_{0i'}$ é a parte do modelo que se modifica em função do ponto de parada no itinerário. Caso $t_{1\to2}$, temos $b_{0i'} = -43.243240$ (menor tempo de espera), se $t_{16\to17}$, $b_{0i'} = 124.07$ (maior tempo de espera)

$$t_{i'k} = \beta_0 + b_{0i'} + \epsilon_{t_{i'k}} \tag{8}$$

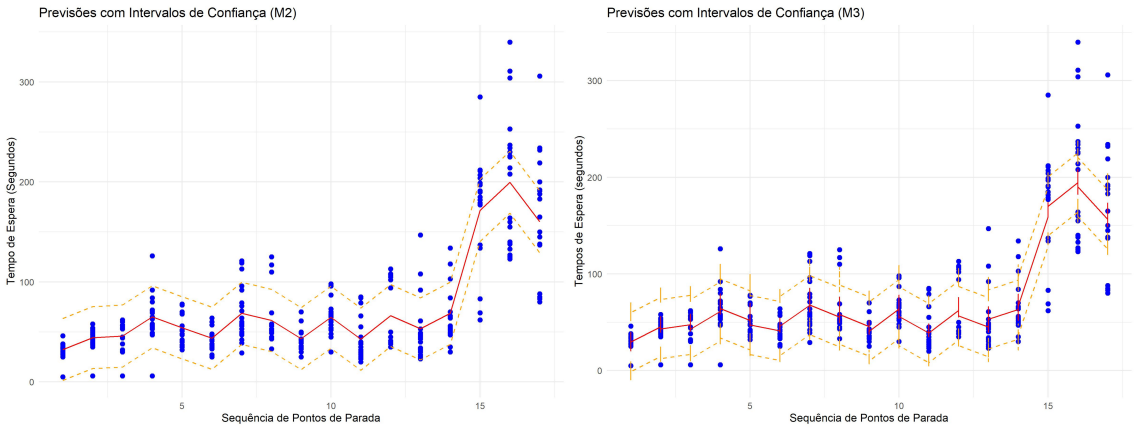

Figura 5. Gráficos de Previsão com Intervalos de Confiança para modelo M2 (Esquerda) e M3 (Direita) considerando um itinerário S.

Tabela 1. EQM Modelos

Modelo		EQM
$M1$		3419.783
$M2$		969.2253
$M3$	$NãoSazonal$	881.3827
$M3$	$Sazonal$	851.4837

4.4. Análise de aderência do modelo aos dados

Para avaliar a aderência do modelo "M3 Sazonal" aos dados reais se aplicou o seguinte método:

1. Treinou-se o modelo "M3 Sazonal" com os dados de uma data arbitrária. A data escolhida foi 11/09/2024 (Quarta-Feira, sem chuva);
2. Escolheu-se uma segunda data arbitrária que tivesse as mesmas características do item 1 (Quarta-Feira, sem chuva). A data escolhida foi 18/09/2024.
3. Plotou-se gráficos para ambas as datas. Através da análise gráfica, procurou-se identificar se os dados do item 2 se encontravam satisfatoriamente dentro do intervalo de confiança definido no item 1.

O resultado do item 3 é exposto na figura 6. Analisando-se a figura se nota um tendência equivalente em ambos os gráficos, no qual a maioria dos pontos se encontra dentro do intervalo de confiança. Os valores fora do intervalo de confiança são poucos em ambas as datas analisadas, não formando uma tendência por si só. Desta forma, é razoável assumi-los como valores atípicos advindos de uma flutuação pontual do comportamento do ônibus/motorista no dia em questão.

5. Considerações Finais

A partir de um estudo de caso com dados históricos de geolocalização de veículos de transporte coletivo percorrendo um itinerário estabelecido pela operação do sistema de transporte público, foram propostos três modelos para buscar representar o comportamento temporal.

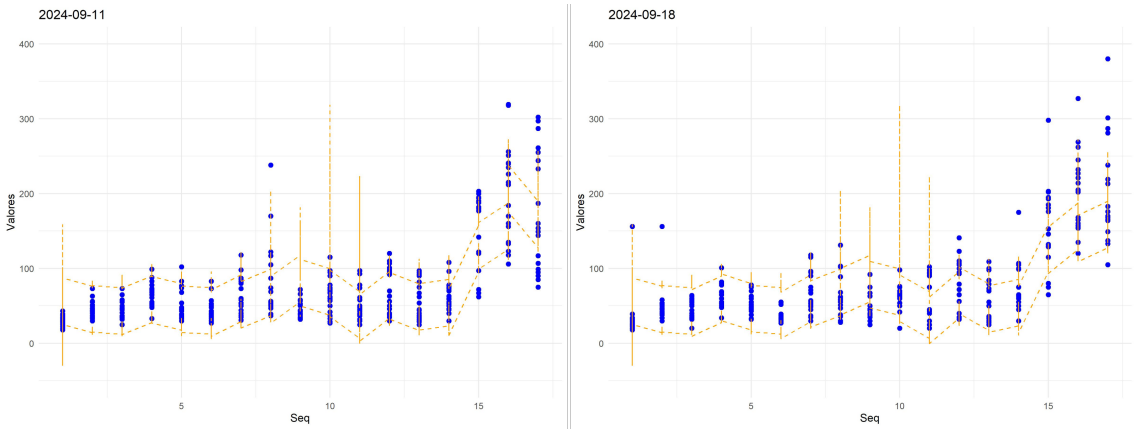

Figura 6. Dados reais e intervalo de confiança para os dias 11/09 e 18/09 de 2024.

O modelo M3 Sazonal apresentou o melhor resultado, em termos de previsão de tempos de espera em pontos de ônibus, bem como na modelagem dos resíduos. Ainda, apresentou-se como uma solução robusta capaz de prever boa parte do comportamento da linha analisada, respeitando o intervalo de confiança estipulado.

Além disso, nota-se que o modelo desenvolvido tem potencial para beneficiar não apenas o usuário final do STPU, disponibilizando tempos de espera e possibilitando a previsão de chegada, mas pode ser utilizado pelo operador do STPU como uma ferramenta de controle de qualidade para identificar comportamentos e atrasos atípicos, permitindo a implementação de alertas e controle proativo.

Como trabalhos futuros e oportunidades de melhoria, se indica a possibilidade de incluir novas variáveis ao modelo, no intuito de melhor explicar as causas de atrasos. Entre as futuras abordagens se destacam: a adequação e uso de dados de Boletins de transporte que informam disrupções programadas pelo operador do STPU na operação da linha do ônibus (devem ser disponibilizados em um formato legível para consumo por computadores); 2. incluir variáveis de possível impacto na operação da linha do ônibus, por exemplo, dados de Pluviometria; de Locais de maior fluxo de passageiros e pontos de integração com outras linhas, por exemplo, ampliando o escopo proposto por [Curzel et al. 2019].

Referências

Borges, J. C., Lüders, R., Silva, T., and Munaretto, A. (2023). Algoritmo para detecção de itinerários do transporte público usando dados de gps dos ônibus. In *Anais do VII Workshop de Computação Urbana*, pages 1–14. SBC.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). Time series analysis: forecasting and control. *John Wiley & Sons*.

Braunschweig, K., Eberius, J., Thiele, M., and Lehner, W. (2012). The state of open data. *Limits of current open data platforms*, 1:72–72.

Büchel, B. and Corman, F. (2022). What do we know when? modeling predictability of transit operations. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15684–15695.

Curzel, J. L., Lüders, R., Fonseca, K. V. O., and Rosa, M. (2019). Temporal performance analysis of bus transportation using link streams. *Mathematical Problems in Engineering*, 2019(1):6139379.

Dong, J., Zou, L., and Zhang, Y. (2013). Mixed model for prediction of bus arrival times. In *2013 IEEE Congress on Evolutionary Computation*, pages 2918–2923.

Faraway, J. J. (2006). Extending the linear model with r: generalized linear, mixed effects and nonparametric regression models. *Chapman and Hall/CRC*.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). Applied longitudinal analysis. *John Wiley & Sons*.

Gallivan, F., Ang-Olson, J., Liban, C. B., and Kusumoto, A. (2011). Cost-effective approaches to reduce greenhouse gas emissions through public transportation in los angeles, california. *Transportation research record*, 2217(1):19–29.

Habib, K. M. N., Kattan, L., and Islam, M. T. (2011). Model of personal attitudes towards transit service quality. *Journal of Advanced Transportation*, 45(4):271–285.

Hashiguchi, K. K., Gai, B. d. F., Pigatto, D. F., and Fonseca, K. V. (2020). Exploratory analysis of public transportation data of curitiba, brazil. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE.

Kumar, B. A., Singh, R., Shaji, H. E., and Vanajakshi, L. (2025). Bus arrival time prediction: A comprehensive review. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–18.

Li, J., Gao, J., Yang, Y., and Wei, H. (2017). Bus arrival time prediction based on mixed model. *China Communications*, 14(5):38–47.

Manika, E. R. (2022). Ambiente para geração de trajetos de linhas de transporte público para análise de mobilidade e viabilidade de comunicação. Master's thesis, Universidade Tecnológica Federal do Paraná.

Martins, T. S. et al. (2022). Map matching: uma análise de dados streaming de trajetórias de gps no transporte público. Master's thesis, Universidade Tecnológica Federal do Paraná.

Matloff, N. (2017). *Statistical regression and classification: from linear models to machine learning*. Chapman and Hall/CRC.

Peixoto, A. M., de Oliveira Rosa, M., Lüders, R., and Fonseca, K. V. O. (2020). Plataforma computacional para construção de um banco de dados de grafo do sistema de transporte de curitiba. In *Anais do IV Workshop de Computação Urbana*, pages 125–137. SBC.

Preston, J. (2009). Transport, public. In Kitchin, R. and Thrift, N., editors, *International Encyclopedia of Human Geography*, pages 452–459. Elsevier, Oxford.

Suwardo, W., Napiah, M., and Kamaruddin, I. (2010). Arima models for bus travel time prediction. *J. Inst. Eng. Malaysia*, 71(2):49–58.

Wimbadi, R. W., Djalante, R., and Mori, A. (2021). Urban experiments with public transport for low carbon mobility transitions in cities: A systematic literature review (1990–2020). *Sustainable Cities and Society*, 72:103023.