UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JOSÉ IVAN SILVA DA CRUZ JÚNIOR

MÉTODOS DE ANÁLISE DE DADOS NO TRANSPORTE PÚBLICO URBANO

CAMPINA GRANDE - PB

2020

JOSÉ IVAN SILVA DA CRUZ JÚNIOR

MÉTODOS DE ANÁLISE DE DADOS NO TRANSPORTE PÚBLICO URBANO

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Claudio Elízio Calazans Campelo.

CAMPINA GRANDE - PB

2020

C957m Cruz Júnior, José Ivan Silva da.
Métodos de análise de dados no transporte público urbano. / José Ivan Silva da Cruz Júnior. – 2020.

11 f.

Orientador: Prof. Dr. Claudio Elízio Calazans Campelo.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Sistema de transporte público. 2. Transporte público urbano. 3. Big data. 4. Métodos de análise de dados. 5. Análise de dados. 6. Análise espacial das origens e destinos das viagens. 7. Análise da lotação dos ônibus por rota. 7. Detecção de outkiers na velocidade de viagens I. Campelo, Claudio Elízio Calazans. II. Título.

CDU:004(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

JOSÉ IVAN SILVA DA CRUZ JÚNIOR

MÉTODOS DE ANÁLISE DE DADOS NO TRANSPORTE PÚBLICO URBANO

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Claudio Elízio Calazans Campelo
Orientador – UASC/CEEI/UFCG

Professor Dr. Carlos Wilson Dantas Almeida
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 2020.

CAMPINA GRANDE - PB

Métodos de análise de dados no transporte público urbano

Trabalho de Conclusão de Curso

José Ivan Silva da Cruz Júnior (Aluno), Cláudio Campelo (Orientador)*
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

RESUMO

Com o crescimento cada vez maior dos centros urbanos e da demanda pelo transporte público, cresceu o interesse por mecanismos que ajudem na análise de dados para melhor compreender a dinâmica do sistema de transporte. Neste artigo, propomos um conjunto de métodos de análise de dados que oferece uma gama de funções úteis para análise dos dados de transporte público, incluindo: a análise espacial das origens e destinos das viagens; a análise da lotação dos ônibus por rota; e a detecção de outliers na velocidade das viagens realizadas. Os métodos foram avaliados através de um estudo de caso na cidade de Curitiba, Brasil, contendo 4169274 viagens percorridas por 246 rotas. Acreditamos que o conhecimento extraído por esse conjunto de métodos de análise e outros similares podem contribuir para melhorar o serviço oferecido a cidadãos em diferentes cidades.

Palavras-chave: *Sistema de Transporte Público. Análise de Dados. Métodos de Análise de Dados. Big Data.*

ABSTRACT

With the increasing growth of urban centers and the demand for public transport, interest in mechanisms that help in data analysis to better understand the dynamics of the transport system has grown. In this article, we propose a set of data analysis methods that offers a range of useful functions for analyzing public transport data, including: spatial analysis of travel origins and destinations; the analysis of bus capacity by route; and the detection of outliers in the speed of the trips made. The methods were evaluated through a case study in the city of Curitiba, Basil, containing 4169274 trips covered by 246 routes. We believe that the knowledge extracted by this set of analysis methods and similar ones can contribute to improving the service offered to citizens in different cities.

Keywords: *Public Transport System. Data Analysis. Data Analysis Methods. Big Data.*

1 INTRODUÇÃO

O Brasil consolida-se como um país notavelmente urbano, com 84,7% de sua população vivendo em áreas urbanas [1]. A busca pelo espaço urbano e, portanto, seu crescimento, demanda da gestão pública constante planejamento, instituição de políticas urbanas, criação de infraestrutura e oferta de serviços públicos ante as necessidades e os interesses cotidianos dos cidadãos, visando o bem-estar de todos [2].

No que tange a oferta de serviços públicos, o transporte público coletivo é tido como um serviço essencial à população [3] e definido como o transporte de passageiros acessível a toda população mediante pagamento individualizado, com itinerários e preços fixados pelo poder público [4].

Nos dias atuais, aproximadamente metade da população mundial vive em cidades. Cerca de 8% vive em cidades com mais de dez milhões de habitantes [5]. Estima-se que o número de pessoas que vive em cidades atingirá cinco bilhões até 2030 [6]. Sendo assim, muitas pessoas precisam usufruir de serviços urbanos que, dada a quantidade de pessoas e tamanho das cidades, são complexos. Todos os dias, com o propósito de trabalhar, estudar ou divertir-se, milhões de pessoas utilizam o transporte público para se deslocar.

Um relatório, divulgado pelo Moovit [7] de 2016 (aplicativo usado diariamente por 350 milhões de passageiros de todo o mundo) sobre o cenário do transporte público no mundo, mostra que aproximadamente um terço dos usuários se desloca por mais de 2 horas diárias em grandes cidades como São Paulo, Cidade do México e Londres. Além disso, quase 40% dos usuários esperam mais de 20 minutos por dia em uma estação de ônibus. O relatório também mostra que, para várias cidades onde o aplicativo é usado, a distância média que as pessoas percorrem em uma única viagem pela cidade vai de 3,6 km em Campina Grande, no Brasil para 11,2 km em Hong Kong, por exemplo.

Nas últimas décadas, a diversidade e quantidade de dados coletados diariamente por nossos sistemas de informação cresceu vertiginosamente. Estima-se que só de 2013 a 2015 foram produzidos dados equivalentes a todos os anos anteriores da história [8]. Tecnologias como Big Data, Computação Ibíqua, Internet das Coisas, entre outras surgiram para facilitar a produção e análise desses dados. Particularmente, no contexto de nossas cidades, hoje temos dados abundantes sobre os locais, as pessoas que habitam nesses locais, vias, trânsito, etc.

Outro fenômeno relevante neste contexto é a recente popularização e evolução de métodos de análise de grandes quantidades de dados, através da Mineração de Dados, ou Ciência de Dados. A aplicação de técnicas de ciência de dados às grandes massas de dados sobre nossas cidades possibilita a cidadãos e gestores compreender como as habitamos e como estas nos afetam em uma escala sem precedentes.

Sendo assim, este trabalho tem como objetivo aplicar técnicas de ciência de dados e visual analytics a dados do transporte público, a fim de propor um conjunto de métodos de análise de dados úteis a pesquisadore e desenvolvedores de software para gestão do

transporte público, visando facilitar a exploração e análise deste tipo de dado por gestores e tomadores de decisão. Para validação da ferramenta proposta, foi conduzido um estudo de caso com dados do transporte público da cidade de Curitiba (Paraná, Brasil).

Diante da diversidade de análises que podem ser implementadas e propostas, saber quais delas serão, de fato, relevantes para os gestores, pesquisadores e usuários do transporte público em geral é um desafio considerável. Diante disso, três análises são implementadas e propostas: análise espacial das origens e destinos finais das viagens; análise da lotação dos ônibus das rotas em qualquer hora do dia; e a detecção de outliers em relação à velocidade das viagens realizadas, com foco da detecção de viagens mais lentas.

O restante deste artigo está estruturado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados à proposta do presente trabalho; a Seção 3 apresenta a abordagem adotada para pré-processamento e preparação da base de dados; a Seção 4 descreve as análises propostas e aplicadas em nosso estudo de caso, mostrando a relevância e aplicabilidade de cada uma; por fim, a Seção 5 mostra as conclusões retiradas das análises, os desafios enfrentados, além de apresentar pontos a serem desenvolvidos em trabalhos futuros.

2 TRABALHOS RELACIONADOS

Com o crescimento cada vez maior das cidades e da necessidade por soluções com alta tecnologia, que otimizem a gestão dos serviços de mobilidade, que auxiliem a tomada de decisões e que proporcionem aos cidadãos uma melhor experiência em suas cidades, tem surgido uma demanda crescente aos pesquisadores, aos cientistas e aos analistas de dados por implementar métodos e ferramentas de análise de dados que possam ser utilizadas por gestores, por pesquisadores e por usuários do transporte público em geral.

Bessa *et al* apresentam o RioBusData [9], uma ferramenta que ajuda os usuários a identificar e explorar, através de diferentes visualizações, o comportamento de trajetórias outlier nos ônibus do sistema de transporte público da cidade do Rio de Janeiro. O sistema desenvolvido detecta automaticamente os outliers usando uma Rede Neural Convolucional. No trabalho, é apresentado uma série de estudos de caso que mostram como a ferramenta ajuda os usuários a compreender melhor não apenas o fluxo e o serviço de ônibus atípicos, mas também o sistema de ônibus como um todo.

Por sua vez, Marques *et al* [10], levando em consideração que muitas soluções de transporte público baseiam-se em informações obtidas de GPS, sejam elas em tempo real, ou armazenadas em uma base de dados fixa, onde o volume de dados dificulta sua análise de maneira manual, propõe uma metodologia para detectar e corrigir erros grosseiros (outliers) em coordenadas geográficas de rotas de linhas de transporte público coletivo, a fim de melhorar a qualidade dos dados disponíveis, visando uma maior precisão nos processos de planejamento de transporte ou mesmo auxiliar na tomada de decisão.

Já o presente trabalho se propõe a analisar os outliers das viagens mais lentas, sendo possível analisá-los pelos dias da semana (segunda a sexta), por um dia da semana específico, uma data específica ou por rota específica observando os dias da semana, buscando oferecer ao gestor uma maior capacidade de perceber onde há a necessidade de serem feitas correções e onde é necessário uma maior investigação para saber as razões que levaram à lentidão de determinadas rotas em determinados horários.

Outra forma de oferecer uma ferramenta relevante a ser usada pelos agentes e pelos pesquisadores do transporte público é reduzir consideravelmente os custos envolvidos nas coletas e tratamento dos dados, como porposto por Galdino (2018) [11], que fez uso de técnicas de Big Data para o carregamento e a matriz origem-destino das linhas de ônibus do transporte público de João Pessoa-PB. No presente trabalho, foi desenvolvido um script na linguagem de programação estatística R, onde as informações de 4169274 viagens percorridas por 246 rotas são processadas para que se torne viável e possível a manipulaçao e análise dessa quantidade de dados.

Por sua vez, Ferreira trabalhou na análise dos padrões de viagens dos idosos no transporte público a partir da análise comportamental de viagens baseada em suas atividades [12]. O processo de análise levou em consideração as características socioeconômicas e de viagens dos idosos do Distrito Federal. A quantidade de viagens observada com maior frequência foi a de duas viagens 68,05%. Já os padrões de viagens, três foram encontrados com maior frequência: HWH (Casa – Trabalho – Casa) com 44,52%, HPH (Casa – Assuntos Pessoais – Casa) com 28,27% e HMH (Casa – Saúde – Casa) com 12,72%. Para a construção dos padrões de viagens levou-se em consideração as viagens que iniciaram e terminaram na residência e o motivo de viagem.

No presente trabalho, uma das análises propostas é a análise espacial de origens e destinos finais das viagens. Onde, apesar de não levar em conta as características socioeconômicas e uma determinada faixa de idade, como foi feito por Ferreira [12], é possível ver quais são os lugares da cidade mais demandados e em quais locais as pessoas mais embarcam em qualquer horário desejado, auxiliando os gestores a determinar quando a oferta de ônibus deverá ser maior para determinados destinos.

3 METODOLOGIA

Os métodos de análise propostos no presente trabalho foram concebidos seguindo os princípios da metodologia KDD (Knowledge-Discovery in Databases)[8], que é um processo de extração de informações de base de dados. Esse método consiste na imersão no domínio da aplicação para compreendê-lo de uma forma mais eficiente.

Para o estudo de caso conduzido, os dados históricos sobre o transporte coletivo de Curitiba são disponibilizados pela administradora local - URBS (Urbanização de Curitiba S/A) em parceria com a Universidade Tecnológica Federal do Paraná - UTFPR. São eles: (i) descrição da organização e de funcionamento do serviço de transporte público da cidade no formato GTFS; (ii) dados de geolocalização dos ônibus (GPS); (iii) registros de bilhetagem eletrônica dos ônibus e dos terminais de Curitiba.

Durante a etapa de processamento e análise dos dados, é utilizada a plataforma de desenvolvimento RStudio. Nela o desenvolvimento se dá principalmente utilizando a linguagem de programação R, que é voltada para análise estatística e criação de visualizações de dados. Adicionalmente, utiliza-se a linguagem de marcação Markdown, uma linguagem simples de marcação que possibilita a transformação das análises em relatórios.

3.1 Dataset

Os dados utilizados na pesquisa foram produzidos por Braz [13], que desenvolveu uma Matriz Origem-Destino a partir dos dados de bilhetagem da cidade de Curitiba-PR[13]. Os dados das viagens utilizados nas análises correspondem ao período de 01/05/2017 até 17/07/2017. No total, a base de dados contém 4169274 viagens e 246 rotas diferentes registradas.

3.2 Pré-processamento dos dados

Essencialmente, um *DataFrame* é uma estrutura de dados bidimensional, composta por linhas e colunas, remetendo a uma planilha. Ele pode ser criado a partir de arquivos, páginas da web ou dados gerados por código. A Figura 1 mostra um exemplo da estrutura de um *DataFrame* com o dataset que inicialmente coletamos para o processamento.

cardNum	user_trip_id	itinerary_id	leg_id	route	busCode
2088629	271297	3	1	625	GA162
2088629	271297	3	3	607	HL328
3811772	271391	6	2	602	GR027
2376410	271392	1	1	535	EA077
2376410	271392	1	3	505	BL320

Figura 1: *DataFrame* inicialmente coletado.

A partir da criação do DataFrame, torna-se possível realizar manipulações nos dados de maneira simples, além de fornecer informações úteis para análises exploratórias a serem realizadas com esses dados. Entretanto, possibilitar que o dataframe possua dados que sejam relevantes e simples de manipular exige um processo de transformação de dados. Para isso, utilizamos o Tidyverse, que é um pacote do R, cuja função é carregar outros pacotes do R como dplyr e o tidyr, para transformarmos a antiga base de dados na base de dados que desejamos. [14].

Logo, o DataFrame inicialmente coletado é submetido a uma etapa de pré-processamento para que resulte em informações objetivas e passíveis de melhor manipulação. Novas variáveis são adicionadas, derivadas a partir de combinações das variáveis originais. Por exemplo, originalmente, na presente base de dados, havia a informação de latitude e logitude do embarque e desembarque da viagem. Após o processamento, a informação de distância percorrida em quilômetros é obtida.

As Figuras 2 e 3 mostram a estrutura do DataFrame antes do processamento e a Figura 4 depois.

cardNum	user_trip_id	itinerary_id	leg_id	route	busCode	tripNum	from_stop_id	start_time
2088629	271297	3	1	625	GA162	10	35713	2017-05-02 16:48:11
2088629	271297	3	3	607	HL328	5	26247	2017-05-02 16:54:03
3811772	271391	6	2	602	GR027	8	25402	2017-05-02 16:16:10
2376410	271392	1	1	535	EA077	14	46874	2017-05-02 16:44:09

Figura 2: *DataFrame* antigo - parte 1

Para o pré-processamento dos dados, é desenvolvido um script na linguagem de programação estatística R. Através dele, obtém-se

from_stop_lat	from_stop_lon	to_stop_id	end_time	to_stop_lat	to_stop_lon	leg_duration
-25.49694	-49.29022	27639	2017-05-02 16:47:07	-25.49211	-49.29321	0 days 00:03:56.
-25.49211	-49.29304	25920	2017-05-02 17:11:07	-25.44031	-49.27180	0 days 00:18:04.
-25.48102	-49.29262	25478	2017-05-02 17:09:17	-25.48968	-49.24926	0 days 00:53:07.
-25.55346	-49.25094	30030	2017-05-02 17:04:48	-25.51710	-49.23028	0 days 00:20:39.
-25.51636	-49.23070	26264	2017-05-02 17:31:09	-25.48171	-49.24697	0 days 00:09:13.

Figura 3: *DataFrame* antigo - parte 2

date	week_day	route	from_time	end_hour	busCode	quantity_trips	duration_median	dist_median	from_hour
2017-05-01	1	10	07:07:06	7	88306	1	18.0	2636.0	7
2017-05-01	1	10	08:48:29	8	88303	1	8.0	1577.0	8
2017-05-01	1	10	09:08:05	9	88306	1	6.0	1781.0	9
2017-05-01	1	10	11:59:02	11	88303	1	21.0	4122.0	11

Figura 4: *DataFrame* processado

as seguintes informações por viagem: *duração*, *quantidade total de viagens*, *distância percorrida*, *velocidade*, *dia da semana* e o *código do ônibus* da viagem feita.

O procedimento de transformação das variáveis consiste nas seguintes etapas:

- Para obtermos a *duração mediana das viagens*, processamos as informações do horário de embarque e desembarque;
- Processamos a informação de cada viagem individualmente para calcular a *quantidade total de viagens*. Cada viagem geralmente é uma linha no DataFrame. Logo, agregando por rota, obtemos a quantidade total;
- A *distância percorrida* (em quilômetros) é calculada a partir das coordenadas (latitude e longitude) do embarque e desembarque e, partir desses dados, é calculada a mediana da distância.
- Para a *velocidade*, usamos a distância percorrida (em quilômetros) e a duração da viagem (em hora);
- A informação do *dia da semana* da viagem feita é obtida através da data e hora.
- O *código do ônibus* onde a viagem foi feita é obtido através do seu identificador.

4 ANÁLISES E RESULTADOS

Esta seção apresenta as análises desenvolvidas nos dados do transporte público de Curitiba-PR, estudo de caso em questão, utilizando os métodos de análise propostos, visando oferecer um conjunto de análises a serem executadas para todos que desejam obter informações relevantes do transporte público em qualquer localidade. Os dados utilizados foram os descritos na Seção 3.

4.1 Análise espacial de origens e destinos finais e dos horários de pico

A primeira análise proposta diz respeito a distribuição da movimentação nos ônibus ao longo de todo o dia. Isto é, o número de viagens que são feitas em cada intervalo de tempo, do início até o fim do dia. Esse intervalo de tempo será definido pelo usuário de acordo com a sua preferência. Assim, ele terá a possibilidade de definir sua própria granularidade temporal (comprimento do intervalo) e poderá trabalhar com os dados aplicados a sua realidade e contexto. No caso deste estudo de caso, para a cidade de Curitiba-PR, o horário

definido foi de 4h às 23h, com intervalos de 1h, conforme pode ser visto na Figura 5.

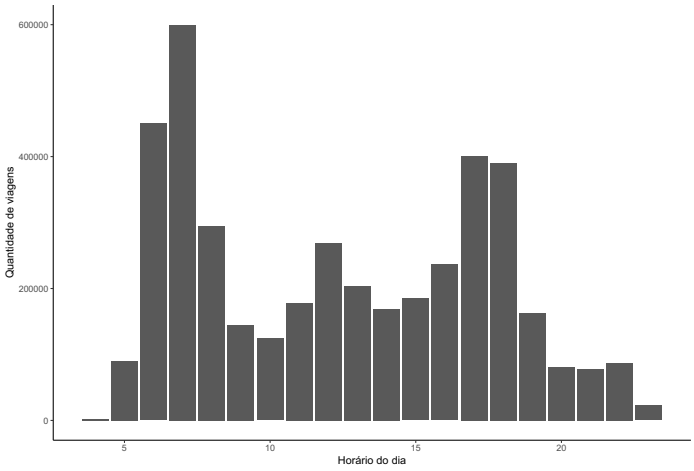

Figura 5: *Quantidade de viagens* durante as horas do dia

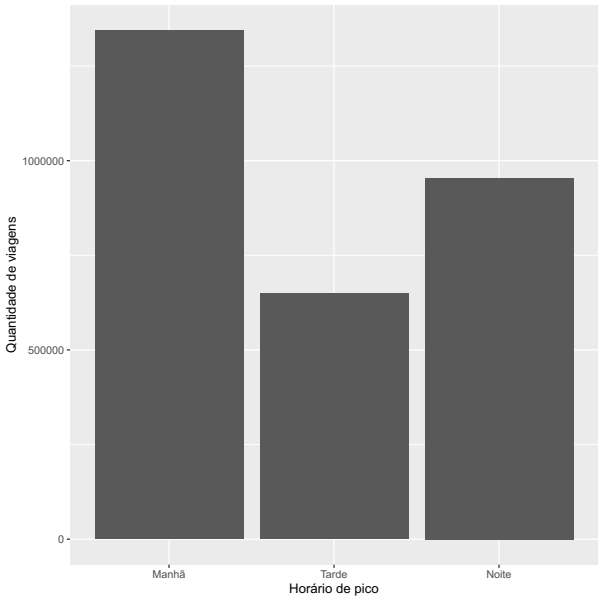

Figura 6: *Horários de pico*

Observa-se também os horários de pico, ou seja, trata-se de saber quais são as faixas de horário em que a movimentação é maior. Os horários de pico demandam uma maior quantidade de viagens e, consequentemente, uma maior quantidade de ônibus e funcionários ativos. A melhor alocação desses recursos é de grande importância para que eles sejam utilizados de forma eficaz, evitando desperdícios e oferecendo o melhor serviço possível.

Como é apresentado na Figura 6, definimos os horários de pico como *manhã*, *tarde* e *noite*, compreendendo os horários de 6h às 8h, 11h às 13h e 17h às 19h, respectivamente. Porém, esse intervalo também é parametrizável, podendo ser definido pelo usuário de acordo com sua necessidade. Isso porque os horários de pico podem ser diferentes de cidade para cidade.

No caso de Curitiba, o horário da *manhã* é o que concentra a maior quantidade de viagens feitas, seguido da *noite* e da *tarde*.

Por fim, é proposta uma análise espacial de origens e destinos finais das viagens. Ou seja, essa análise possibilita a visão de quais são os lugares da cidade mais demandados e em quais locais as pessoas mais embarcam. Essa análise também é feita de acordo com os horários escolhidos pelo pesquisador, gestor ou usuário. Assim, o usuário do método proposto pode analisar a distribuição das origens e destinos de acordo com o seu interesse, incluindo os horários de pico ou não.

Vale ressaltar que a análise é a de origens e destinos finais do passageiro. Isto é, todas as viagens intermediárias que um passageiro faz até chegar seu destino, passando por diferentes pontos ou terminais de ônibus, não são considerados. Assim, é possível observar, onde, de fato, os passageiros embarcaram em seu destino inicial e onde desembarcaram para o seu destino final.

Para a faixa de horário com mais viagens em Curitiba, das 6h às 8h, como mostra a Figura 7, o embarque mostra-se bem distribuído em bairros periféricos e distantes do centro, onde geralmente se encontram bairros residenciais. Constata-se também, pela Figura 8, que, no mesmo horário, o desembarque se mostra mais acentuado e concentrado na região central da cidade. Por outro lado, na faixa de horário das 17h às 19h, o desembarque se acentua na região periférica da cidade (Figura 9) e, o embarque, na região central (Figura 10). Isso indica que na faixa de horário da manhã as pessoas tendem a sair de seus bairros em direção a região central e bairros comerciais da cidade para trabalhar, estudar, etc, e, a noite, voltam para as suas casas.

Observar a distribuição do destino das viagens nos mostra o grau da demanda de ônibus para determinadas regiões da cidade. Logo, saber que há muitas pessoas desejando ir para determinadas regiões, em algum horário do dia, auxiliará os gestores a determinar quando a oferta de ônibus deverá ser maior para os destinos mostrados. Por exemplo, as aulas no campus da Universidade Federal do Paraná em Curitiba se iniciam às 8h, logo a demanda de ônibus para o Campus antes das 8h será intenso em qualquer terminal da cidade. O método de análise permite a análise das demandas em qualquer outra faixa de horário.

4.2 Lotação dos ônibus por dia, horário e rota

Uma realidade do transporte público é a sua propensão a atingir o limite da lotação em determinados horários e itinerários. Em muitos momentos, devido a isso, os usuários demoram mais tempo do que planejaram para fazer uma viagem e os gestores precisam lidar com maiores desafios de gestão em seus sistemas de transporte.

Uma outra análise proposta pelo presente trabalho é prover um panorama sobre a lotação dos ônibus seja qual for a linha, o horário ou dia pretendido. Sendo assim, o usuário da ferramenta pode verificar a lotação dos ônibus de uma determinada rota em qualquer dia ou horário que for do seu interesse, sendo possível verificar a quantidade de passageiros em cada ônibus que esteve fazendo viagem para uma linha específica em qualquer horário pretendido.

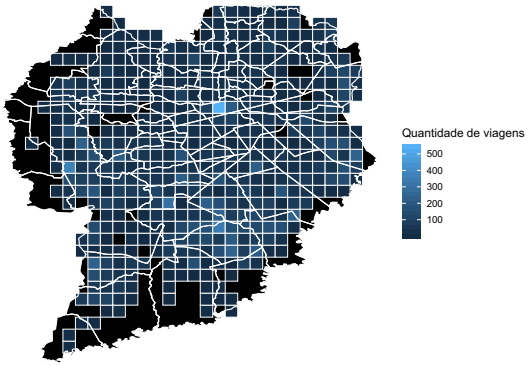

Figura 7: *Embarque* das viagens das 6:00 às 8:00

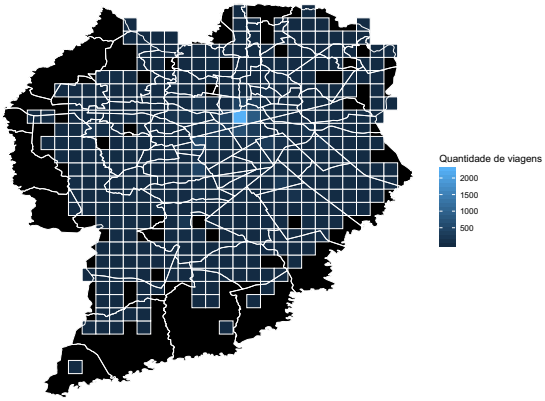

Figura 8: *Desembarque* das viagens das 6:00 às 8:00

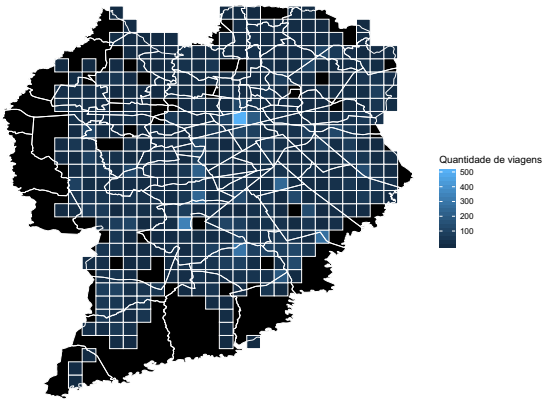

Figura 9: *Desembarque* das viagens das 17:00 às 19:00

Dessa forma, os gestores terão uma forma mais fácil de fiscalizar a lotação por horário, obtendo maior capacidade de planejar e dispor os recursos demandados para o oferecimento de um serviço

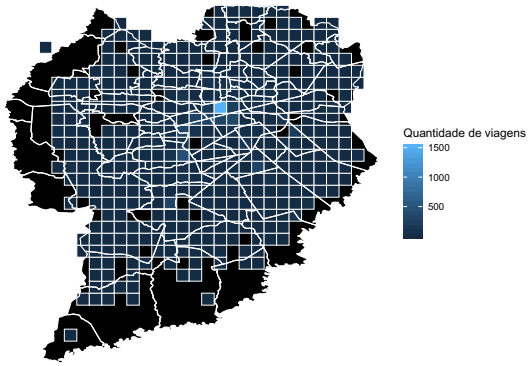

Figura 10: *Embarque* das viagens das 17:00 às 19:00

de melhor qualidade. Ao observar um desequilíbrio na quantidade de passageiros para os ônibus de uma mesma linha no horário, ele poderá procurar soluções para uma melhor distribuição de passageiros nos ônibus, evitando uma desigualdade relevante na quantidade de passageiros transportados para o mesmo destino na mesma faixa de horário.

Ressalta-se que a análise não é realizada indicando a quantidade de passageiros viajando naquele exato momento, mas sim aqueles que fizeram check-in na rota no horário especificado.

Para Curitiba, em um exemplo de aplicação da análise para a rota 303 na faixa de horário de 18h às 19:59h do dia 02/05/2017, observa-se na Figura 11 que, no geral, os ônibus seguem uma média de quantidade de passageiros parecida no decorrer do horário, com a exceção de dois ônibus que apresentam uma quantidade de passageiros transportados bem maior que os demais, chegando a ter mais de 230 passageiros embarcando nos ônibus durante a faixa de horário.

O gestor, ao observar a distribuição temporal dos ônibus e a quantidade de passageiros transportados, poderá enxergar através da transição de um ônibus para outro o desequilíbrio que pode haver entre a lotação dos dois. É importante destacar que os ônibus estão ordenados pelo primeiro check-in realizado na linha dentro do horário especificado.

4.3 Detecção de outliers (viagens lentas)

Os outliers são dados que se diferenciam drasticamente de todos os outros, são pontos fora da curva. Em outras palavras, um outlier é um valor que foge da normalidade e que pode causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise.

Detectar outliers se apresenta como uma tarefa de extrema importância para descoberta de conhecimento e mineração de dados, especialmente ao lidarmos com problemas em ciências aplicadas. E, embora seja um tema abordado a um razoável tempo através de métodos estatísticos, se renova como um tópico de pesquisa de grande relevância nos dias atuais, devido ao grande crescimento na disponibilidade de dados para os pesquisadores e indústria. Em diversos cenários, os dados são tantos que realizar o processamento de todo o conjunto disponível é impraticável ou até mesmo indesejável. Assim, métodos capazes de selecionar aqueles dados com alto

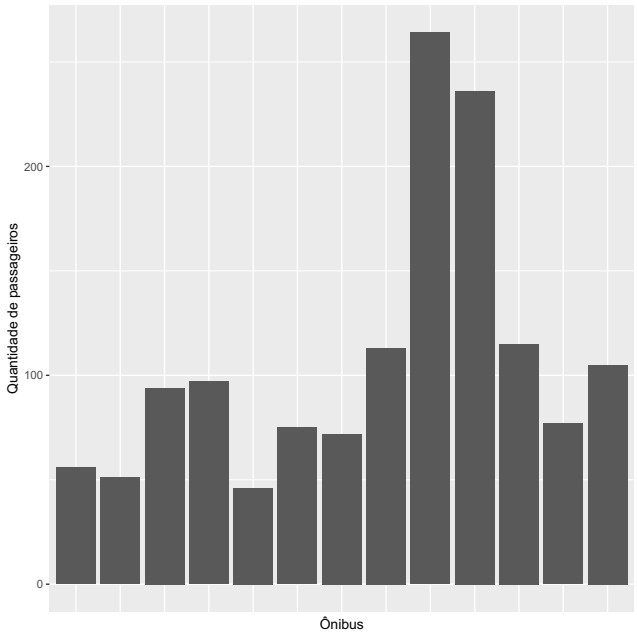

Figura 11: *Quantidade de passageiros* transportados nos ônibus na linha 303 de 18:00min às 19:59min do dia 02/05/2017

grau de distinção em meio a todo esse volume despertam grande interesse. [15]

Entender os outliers é fundamental em uma análise de dados por pelo menos dois aspectos:

- os outliers podem viesar negativamente todo o resultado de uma análise;
- o comportamento dos outliers pode ser justamente o que está sendo procurado.

Os outliers possuem diversos outros nomes, como: dados discrepantes, pontos fora da curva, observações fora do comum, anomalias, valores atípicos, entre outros. [16]

Diante disso, o presente trabalho sugere uma análise para detecção de outliers objetivando buscar as viagens mais lentas realizadas. A intenção é dispor também ao usuário uma possibilidade de pré-processamento dos dados em função daqueles que fogem do comportamento esperado.

Nos dados de Curitiba-PR, podemos ver na Figura 12 que o padrão encontrado na relação entre distância percorrida e duração da viagem é diretamente proporcional. Isso quer dizer que, quanto maior for a distância da viagem, mais tempo o usuário demorará para chegar a seu destino. Nosso objetivo é analisar as viagens que fogem desse padrão e se mostram como mais lentas que o normal. O critério utilizado para definir uma viagem como lenta foi a sua distância da nuvem de dados da Figura 12 que concentra a maior quantidade de viagens.

O método de análise concede ao usuário as seguintes possibilidades de detecção dos outliers:

Vale ressaltar que os resultados mostrados nas figuras são da cidade de Curitiba-PR, estudo de caso do trabalho.

- Observar a quantidade de viagens lentas por dia da semana, mostrando assim quais os dias da semana onde as viagens tendem a ser mais lentas (Figura 13);
- Escolher um dia da semana específico e observar quais rotas apresentam a maior quantidade de viagens lentas. Nesse estudo de caso, o dia escolhido foi a quinta-feira, porém a análise permite a escolha de qualquer dia da semana (Figura 14);
- Escolher uma data específica e ver quais rotas apresentaram a maior quantidade de viagens lentas. A data escolhida foi 10/05/2017(Figura 15);
- Escolher uma rota específica e observar a quantidade de viagens lentas nos dias da semana. A rota escolhida foi a 370 (Figura 16).

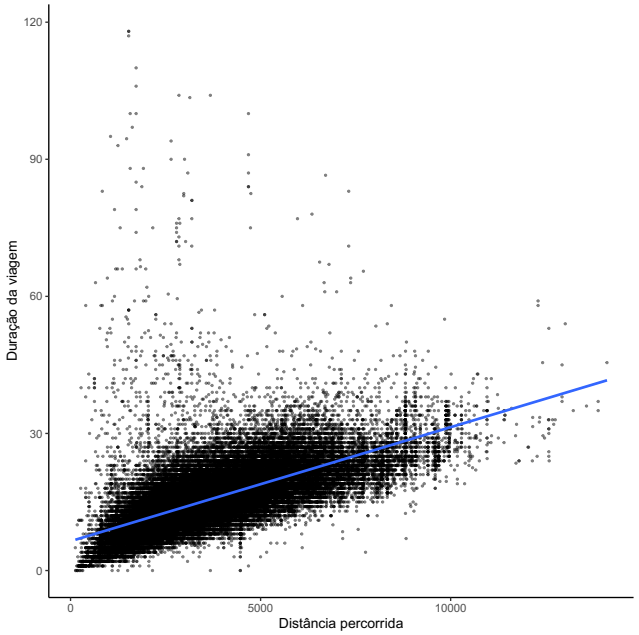

Figura 12: *Relação* da distância percorrida (km) e duração das viagens (minutos) de todos os dados da base de viagens.

A possibilidade de detectar as viagens mais lentas no transporte público, nas diversas formas apresentadas acima, concede ao gestor uma maior capacidade de perceber onde há a necessidade de serem feitas correções e onde é necessário uma maior investigação para saber as razões que levaram à lentidão de determinadas linhas em determinados horários. A análise proposta é útil também para entender o funcionamento do transporte nos dias atípicos, ou seja, aqueles onde grandes eventos são realizados na cidade, exigindo uma dinâmica diferente da oferta de ônibus e da demanda de viagens em um horário específico.

5 CONCLUSÃO

Para a implementação das análises, foi escolhida a linguagem de programação R tanto pela grande popularidade no desenvolvimento de análises, manipulação, e visualização de dados como pela usabilidade, facilidade na codificação, legibilidade e reutilização em

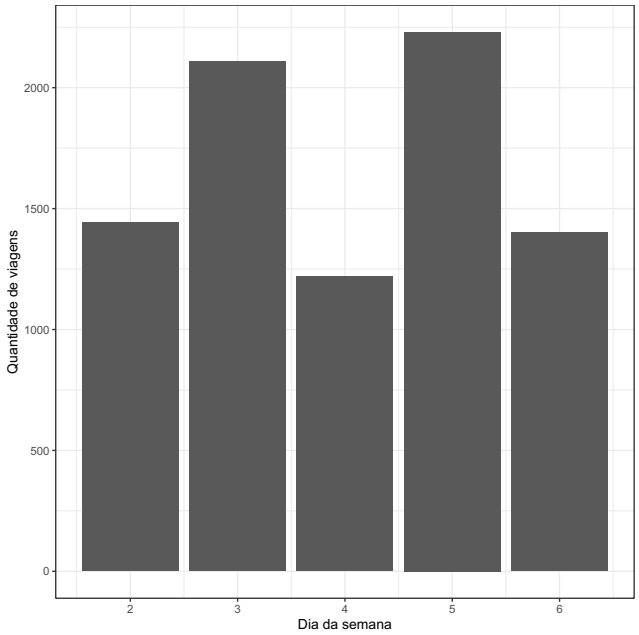

Figura 13: *Quantidade* de viagens lentas por dia da semana.

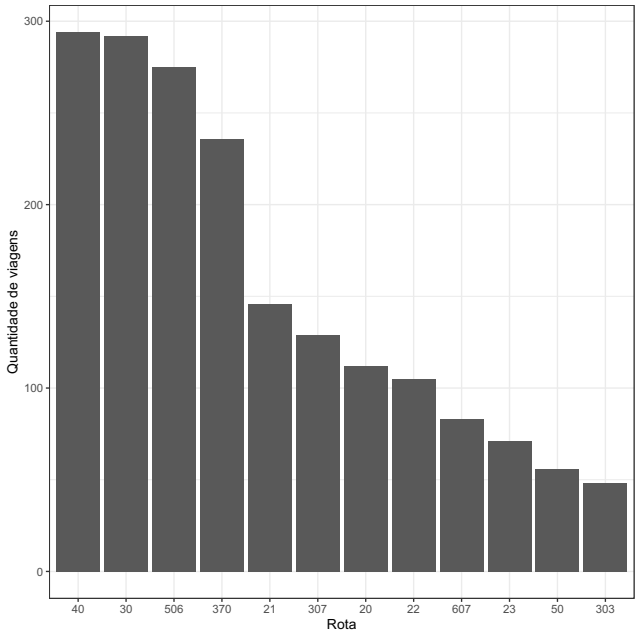

Figura 14: *Quantidade* de viagens lentas por rota (quinta-feira)

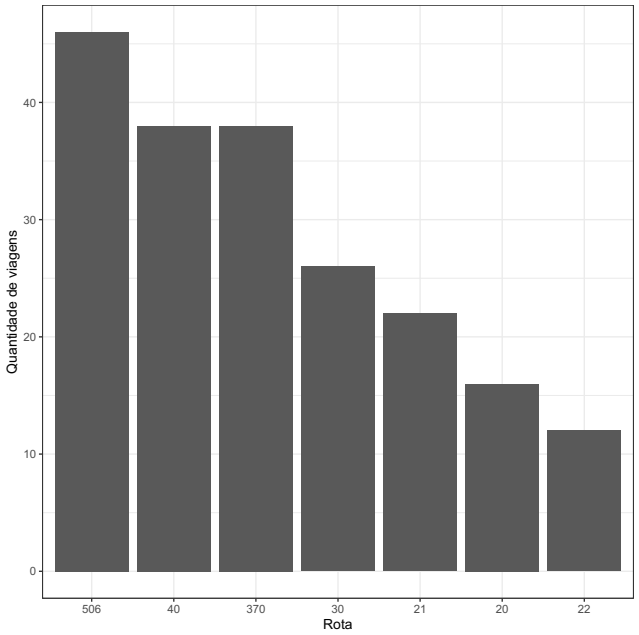

Figura 15: *Quantidade* de viagens lentas por rota através da data (10/05/2017)

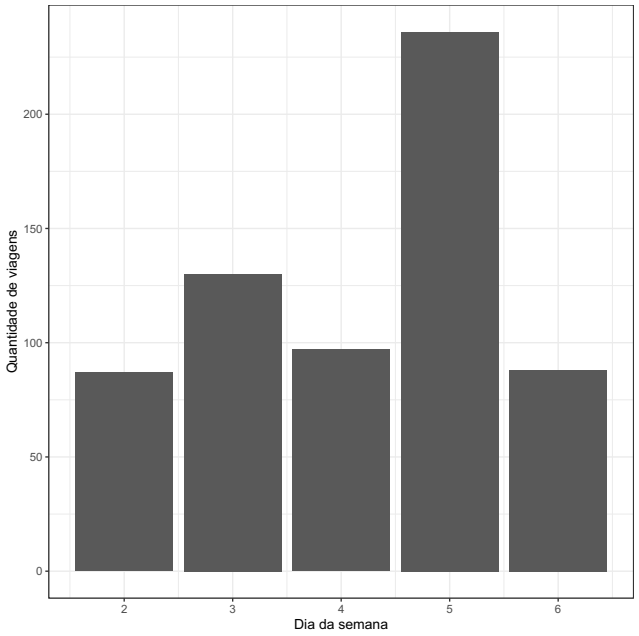

Figura 16: *Quantidade* de viagens lentas por dia da semana através da rota (370)

projetos de análises de dados. O desenvolvimento se deu pela escolha de três análises que pudessem ser aplicadas em qualquer contexto de pesquisa no transporte público.

O maior desafio encontrado foi conceber e desenvolver análises que fossem realmente relevantes para os gestores e aqueles que trabalham com pesquisa no transporte público, em vistas da base de dados disponível. Pensar na utilidade das análises exigiu reuniões e refinamentos constantes com o orientador objetivando o

esclarecimento de quais poderiam, de fato, agregar e ser útil para aqueles que buscam por informações nesse contexto.

Em relação a aprimoramentos que podem ser realizados e novas análises que podem ser propostas em trabalhos futuros, podem ser listados:

- Expansão da detecção dos outliers para a demanda de locais de destino;
- Visualizações mais avançadas para as análises espaciais;
- Visualizações mais avançadas para as análises de lotação dos ônibus;
- Análise da lotação dos ônibus em tempo real.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus por me conduzir e me capacitar em toda graduação. A Ele o meu tributo.

Aos meus pais José Ivan e Aparecida, por toda abnegação e esforço que dedicaram na minha educação.

Ao meu tio Ednaldo por ter me acolhido em sua casa durante a graduação e a minha tia Magna por todo suporte dado nesse tempo.

Ao meu grande amigo Tarciso, pelo companheirismo e encorajamento dados na minha caminhada na universidade e pelas contribuições valiosas para que esse trabalho fosse finalizado.

Ao professor Campelo por ter sido um excelente orientador e pelas essenciais contribuições dadas sempre com muita humanidade e paciência.

A Missão Federal, por ter sido refúgio e suporte na caminhada do curso.

Ao povo brasileiro, por proporcionar educação superior pública de qualidade.

Por fim, agradeço a todos os meus amigos que me deram forças nessa jornada.

REFERÊNCIAS

[1] Conheça o brasil - população.
[2] Lei n. 10.257, de 10 de julho de 2001.
[3] Constituição (1988). constituição da república federativa do brasil.
[4] Lei n. 12587, de 03 de janeiro de 2012.
[5] Mais de metade da população mundial já vive em áreas urbanas.
[6] Mundo terá 9 bilhões de pessoas em 2050, diz onu.
[7] Índice do moovit sobre o transporte público.
[8] GREGORY PIATETSKY-SHAPIRO USAMA M. FAYYAD and PADHRAIC SMYTH. "from data mining to knowledge discovery: an overview". pages 1–34, 2020.
[9] Aline Bessa. Riobusdata: Outlier detection in bus routes of rio de janeiro. *arXiv preprint arXiv:1601.06128*, 2016.
[10] Rodney R. Saldanha; Ramon L. Marques; Sérgio L. Sermenho. Avaliação de métodos para detecção e correção de outliers em coordenadas geográficas em linhas de transporte público. 2013.
[11] Diego Ribeiro de Oliveira Galdino. Análise de dados de gps e bilhetagem eletrônica para determinação do carregamento e matriz de origem-destino no sistema de transporte público por ônibus de joão pessoa. 2018.
[12] FERREIRA Noemia Gomes. Análise dos padrões de viagens do idoso em relação ao transporte público. *Universidade de Brasília*, 2012.
[13] BRAZ T. Inferring passenger-level bus trip traces fromschedule, positioning and ticketing data: Methodsand applications. *Universidade Federal de Campina Grande (UFCG)*, 2019.
[14] Princípios de uso do r.
[15] Rafael Delalibera Rodrigues. Detecção de outliersbaseada em caminhada determinística do turista. *Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto/USP*, 2018.
[16] Ouliers, o que são outliers e como trata-los em uma analise de dados?