MobilityHelp: Uma Ferramenta para Análise de Dados no Transporte Público Urbano

José Ivan S. da Cruz Júnior[1], **Claudio E. C. Campelo**[1]

[1]Universidade Federal de Campina Grande (UFCG)
Departamento de Sistemas e Computação
Campina Grande – PB – Brasil

jose.ivan.junior@ccc.ufcg.edu.br, campelo@dsc.ufcg.edu.br

Abstract. *With the increasing growth of urban centres and the demand for public transport, there has been a considerable increase in the demand for tools that help analyse data to better understand the dynamics of the transport system. In this article, we propose MobilityHelp, a software tool, implemented in the R language, which offers a set of useful functions for analysing public transport data, including: spatial analysis of travel origins and destinations; analysis of bus capacity by route; and detection of outliers in the speed of the trips made. The tool has been evaluated through a case study in the city of Curitiba, Brazil. We believe the knowledge extracted by this and other similar tools can contribute to the improvement of the service offered to citizens in different cities.*

Resumo. *Com o crescimento cada vez maior dos centros urbanos e da demanda pelo transporte público, cresceu o interesse por ferramentas que ajudem na análise de dados para melhor compreender da dinâmica do sistema de transporte. Neste artigo, propomos a MobilityHelp, uma ferramenta de software, implementada na linguagem R, que oferece um conjunto de funções úteis para análise de 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. A ferramenta foi avaliada através de um estudo de caso na cidade de Curitiba, Basil. Acreditamos que o conhecimento extraído por esta ferramenta e outras similares pode contribuir para melhorar o serviço oferecido a cidadãos em diferentes cidades.*

1. Introdução

Este trabalho tem como objetivo aplicar técnicas de ciência de dados e visual analytics a dados do transporte público, a fim de desenvolver um ferramental útil a desenvolvedores de software para gestão de 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 de 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 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.

2. Metodologia

A metodologia adotada para analisar os dados é o KDD (Knowledge-Discovery in Databases)[FAYYAD 2020], 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.

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

Os dados utilizados no estudo de caso foram produzidos por Braz [BRAZ 2019], que desenvolveu uma Matriz Origem-Destino a partir dos dados de bilhetagem da cidade de Curitiba-PR [BRAZ 2019]. 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.

Para o pré-processamento dos dados, é desenvolvido um script na linguagem de programação estatística R. Através dele, obtém-se as seguintes informações por viagem: *duração*, *quantidade total de viagens*, *distância percorrida*, *velocidade*, *dia da semana* e o *código do ônibus* da viagem realizada.

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 era uma linha no dataframe. Logo, agregando por rota, obtivemos a quantidade total;
- A *distância percorrida* (em quilômetros) foi calculada a partir das coordenadas (latitude e longitude) do embarque e desembarque e, a partir desses dados, foi calculada a mediana da distância.
- Para a *velocidade*, usamos a distância percorrida (em quilômetros) e a duração da viagem (em hora);
- O *código do ônibus* onde a viagem foi realizada foi obtido através do seu identificador.

A etapa de pré-processamento dos dados inclui ainda uma atividade de filtragem. onde rotas com pouca quantidade de viagens por dia são excluídas. Para o estudo de caso conduzido, foram excluídas as rotas com menos de 10 viagens por dia. Este limiar pode ser ajustado para outras análises.

3. Análises e Resultados

3.1. Análise espacial de origens e destinos finais

A primeira análise proposta diz respeito a distribuição das 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 é realizada de

acordo com os horários escolhidos pelo pesquisador, gestor ou usuário. Assim, o usuário da ferramenta pode analisar a distribuição das origens e destinos de acordo com o dia e horário do seu interesse. 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 1, o embarque mostra-se bem distribuído em bairros periféricos e distantes do centro, onde geralmente se encontram bairros residenciais. Constata-se também, pela Figura 2, 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 4) e, o embarque, na região central (Figura 3). 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 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.

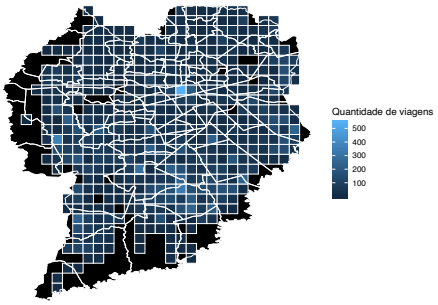

Figure 1. *Embarque* das viagens das 6:00 às 8:00

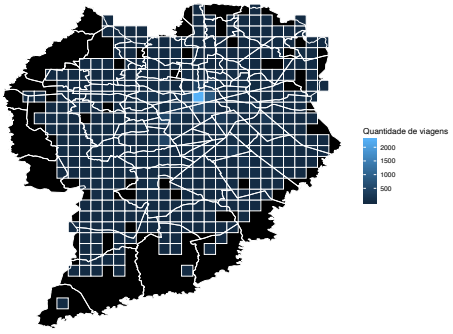

Figure 2. *Desembarque* das viagens das 6:00 às 8:00

3.2. 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. Ressalta-se que a análise não é realizada indicando a quantidade

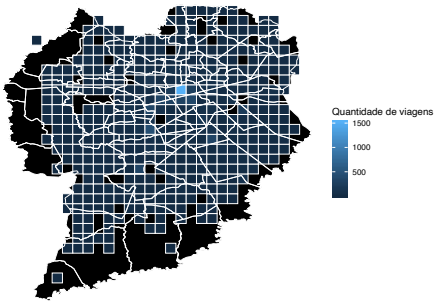

Figure 3. *Embarque* das viagens das 17:00 às 19:00

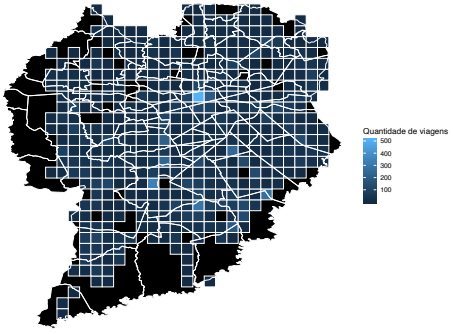

Figure 4. *Desembarque* das viagens das 17:00 às 19:00

de passageiros viajando naquele exato momento, mas sim aqueles que fizeram check-in na rota no horário espefificado.

Para Curitiba, em um exemplo de aplicação da ferramenta para a rota 303 na faixa de horário de 18h às 19:59min do dia 02/05/2017, observa-se na Figura 5 que, no geral, os ônibus seguem uma média de quantidade de passageiros parecida no decorrer do horário, com a excessão 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.

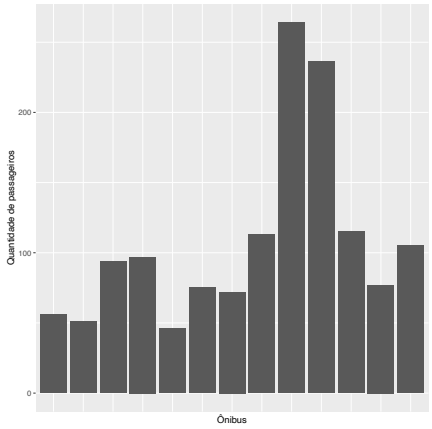

Figure 5. *Quantidade de passageiros* transportados nos ônibus na linha 303 de 18:00min às 19:59min do dia 02/05/2017

O gestor, ao observar a distribuição temporal dos ônibus e a quantidade de passageiros transportados, terá uma forma mais fácil de fiscalizar a lotação por horário, obtendo maior capacidade de planejar e dispor os resursos demandados para o oferecimento de um serviço de melhor qualidade.

3.3. Detecção de outliers (viagens lentas)

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. 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 grau de distinção em meio a todo esse volume despertam grande interesse. [RODRIGUES 2018]

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

Nos dados de Curitiba-PR, podemos ver na Figura 6 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 e, consequentemente, as rotas de ônibus que sistematicamente fogem desse padrão e apresentam velocidades baixas. Viagens e rotas de ônibus que são sistematicamente lentas podem ser alvo de ações estruturantes ou análises mais profundas quanto ao itinerário. O critério utilizado para definir uma viagem como lenta foi a sua distância da nuvem de dados da Figura 6 que concentra a maior quantidade de viagens.

A ferramenta concede ao usuário as seguintes possibilidades de detecção dos outliers:

- 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 7);
- Escolher uma data específica e ver quais rotas apresentaram a maior quantidade de viagens lentas. A data escolhida foi 10/05/2017(Figura 8);
- Escolher uma rota específica e observar a quantidade de viagens lentas nos dias da semana. A rota escolhida foi a 370 (Figura 9).

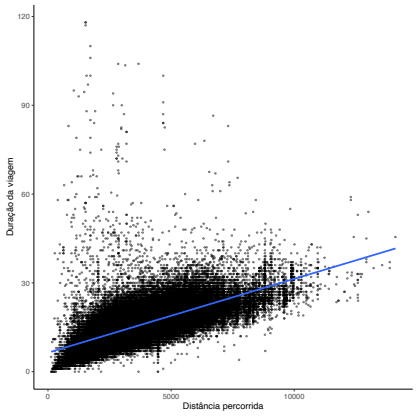

Figure 6. *Relação* da distância percorrida (km) e duração das viagens (minutos) de todos os dados da base de viagens.

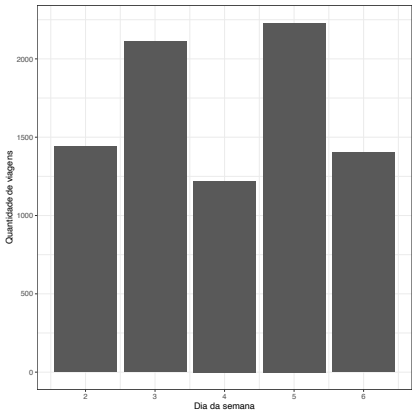

Figure 7. *Quantidade* de viagens lentas por dia da semana.

A ferramenta concede ao usuário uma maior capacidade de perceber onde há a necessidade de serem feitas correções, além de entender o funcionamento do transporte

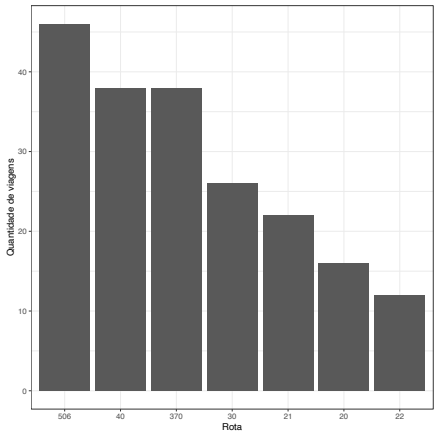

Figure 8. *Quantidade* de viagens lentas por rota através da data (10/05/2017)

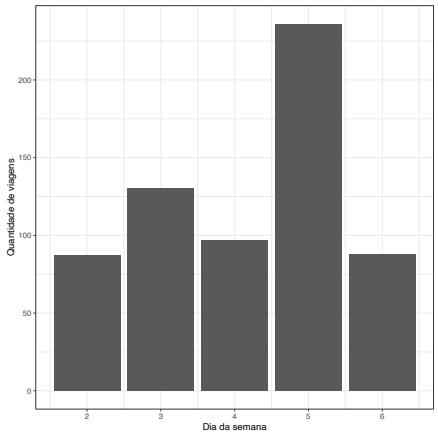

Figure 9. *Quantidade* de viagens lentas por dia da semana através da rota (370)

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.

4. Conclusão

Para a implementação da ferramenta, foi escolhida a linguagem de programação R tanto pela grande popularidade no desenvolvimento de análises, manipulação, quanto facilidade na codificação, legibilidade e reutilização em 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. Possíveis aprimoramentos da ferramenta incluem: expansão da detecção dos outliers para a demanda de locais de destino; visualizações mais avançadas para as análises espaciais; visualizações elaboradas para as análises de lotação dos ônibus; e análise da lotação dos ônibus em tempo real.

References

BRAZ, T. (2019). Inferring passenger-level bus trip traces fromschedule, positioning and ticketing data: Methodsand applications. *Universidade Federal de Campina Grande (UFCG)*.

FAYYAD, U. M. (2020). “from data mining to knowledge discovery: an overview”. pages 1–34.

RODRIGUES, R. D. (2018). Detecção de outliersbaseada em caminhada determinística do turista. *Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto/USP*.