

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA

MÁRIO ALBERTO CECCHI RADUAN

**Análises do transporte público
rodoviário no Rio de Janeiro através
de Dados Abertos**

Profa. Valeria Menezes Bastos, D.Sc.
Orientadora

Rio de Janeiro, Novembro de 2016

Análises do transporte público rodoviário no Rio de Janeiro através de Dados Abertos

Mário Alberto Cecchi Raduan

Projeto Final de Curso submetido ao Departamento de Ciência da Computação do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários para obtenção do grau de Bacharel em Informática.

Apresentado por:

Mário Alberto Cecchi Raduan

Aprovado por:

Profa. Valeria Menezes Bastos, D.Sc.

Profa. Adriana Santarosa Vivacqua, D.Sc.

Prof. Vinícius Gusmão Pereira de Sá, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

Novembro de 2016

Agradecimentos

Ao longo da minha graduação, tive o privilégio de conhecer e trabalhar com diversas pessoas que foram essenciais para que eu chegasse até aqui.

Apesar das dificuldades passadas pela Universidade durante esses anos, incluindo duas longas greves, pude contar com a ajuda de excelentes profissionais que me ajudaram ao longo desse caminho. Entre eles, agradeço imensamente ao mestre Adriano J. O. Cruz, por ter me apresentado ao mundo da pesquisa através de duas iniciações científicas - que me envolveram muito mais na Universidade do que eu podia esperar - e por ter me ensinado tanto de Computação. Agradeço também à professora Valeria Menezes Bastos, por ter sido uma excelente professora e orientadora, e que me motivou tanto durante a reta final da minha graduação.

Agradeço a todo o pessoal do LabIC - Laboratório de Inteligência Computacional - que me acolheu durante a maior parte da graduação, não só como um laboratório, como também um refúgio pessoal e um lugar de muitas amizades.

Agradeço à minha família - meu pai, José Raduan, minha mãe, Ana Elisa Cecchi, minha irmã, Daniela Cecchi, meu irmão, Giancarlo Raduan e meu avô, Mario Fernando Cecchi - por todo o apoio dado durante esses anos. Vocês foram incríveis e me motivaram para que eu chegasse até aqui, sempre pensando na minha educação, minha felicidade e meu futuro.

Agradeço a todo o pessoal do Rio Bus, especialmente ao Fred Souza e Marquinho Tormenta, com quem aprendi muito e serviu como inspiração para este projeto final.

Agradeço aos meus amigos - da turma do BCC de 2011.1, do LabIC, de Glasgow

e da Ilha - por terem me mantido são durante a faculdade, mesmo nas horas em que isso parecia impossível. Agradeço também ao Danilo Vettorazzi, que além de ter sido um grande amigo durante toda a faculdade, me ajudou a escolher o tema deste projeto final e pediu para que eu colocasse o nome dele aqui.

Por fim, agradeço ao meu namorado, Marcos Moretto, por ter sido um ótimo companheiro e, principalmente, amigo durante os últimos anos. Você foi essencial em todos os aspectos.

RESUMO

Análises do transporte público rodoviário no Rio de Janeiro através de Dados

Abertos

Mário Alberto Cecchi Raduan

Novembro/2016

Orientadora: Valeria Menezes Bastos, D.Sc.

Em tempos onde diversos dados governamentais passam a ser disponibilizados de maneira transparente para o público, começam a surgir diversas soluções de terceiros que fazem uso desses dados para auxiliar a população. Esses, agora processados e analisados, tornam-se ferramentas de informação nas mãos do cidadão e promovem a fiscalização dos serviços públicos prestados.

O escopo deste trabalho é baseado nos dados abertos da frota de ônibus da cidade do Rio de Janeiro, um dado fornecido pela Prefeitura carioca e que diz respeito às linhas municipais e à geolocalização de todos os ônibus em tempo real. Utilizaremos esses dados para elaborar algoritmos que permitem analisar o real funcionamento dos ônibus do Rio. Tais algoritmos colhem estatísticas que podem servir como um documento para a fiscalização desse serviço público tão essencial.

ABSTRACT

Análises do transporte público rodoviário no Rio de Janeiro através de Dados

Abertos

Mário Alberto Cecchi Raduan

Novembro/2016

Advisor: Valeria Menezes Bastos, D.Sc.

In times where all kinds of governamental data become available in a transparent manner to the public, different third-party solutions arise making use of this data to help people. That data, now processed and analyzed, become information tools on the citizen's hands and promote the fiscalization of public services.

The studies presented in this document are based on the open data of Rio de Janeiro's city buses, made available by Rio's city hall, composed of information about the municipal bus lines and the geolocalization of all of their buses in real-time. We'll gather data to elaborate algorithms that allow the analysis of the real functioning of Rio's buses. Those algorithms gather statistics that can be used to promote the accountability of such an essential public service.

Lista de Figuras

Figura 2.1: Classificação dos dados abertos segundo a escala de 5 estrelas de Berners-Lee	6
Figura 4.1: Análise mostra o tempo médio de espera por um ônibus da linha 324 num dia útil em julho de 2015.	17
Figura 4.2: Análise mostra o tempo médio de espera por um ônibus da linha 324 num domingo em julho de 2015.	17
Figura 4.3: Vista geral da cidade mostra uma grande concentração de ônibus nas regiões centrais. (Outubro de 2015)	19
Figura 4.4: Comparação da quantidade de ônibus na Zona Sul em dois anos consecutivos. Em 2016, observa-se a redução em virtude da racionalização das linhas de ônibus.	21
Figura 4.5: No Centro, concentração alta nas vias principais e baixa nas transversais. (Outubro de 2015)	22

Lista de Tabelas

Tabela 3.1: Exemplo da coleção histórica de um ônibus 11

Lista de Abreviaturas e Siglas

API	Application Programming Interface
CSV	Comma-separated values
DAG	Dados abertos governamentais
FETRANSPOR	Federação das Empresas de Transporte de Passageiros do Estado do Rio de Janeiro
GPS	Global Positioning System
JSON	JavaScript Object Notation
OD	Open Definition
UFRJ	Universidade Federal do Rio de Janeiro
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

Sumário

Agradecimentos	i
Resumo	iii
Abstract	iv
Lista de Figuras	v
Lista de Tabelas	vi
Lista de Abreviaturas e Siglas	vii
1 Introdução	1
1.1 Objetivos	2
1.2 Motivação	2
1.3 Trabalhos relacionados	3
1.4 Estrutura e metodologia	4
2 Conceitos	5
3 Rio Bus	9

3.1	Base de dados	10
3.1.1	Origem	10
3.1.2	Formatos	11
3.1.3	Armazenamento	11
4	A análise dos dados	13
4.1	Frequência dos ônibus	14
4.1.1	Dados utilizados	14
4.1.2	O algoritmo	14
4.1.3	Implementação	16
4.1.4	Resultados observados	16
4.2	Concentração da frota	19
4.2.1	Implementação	19
4.2.2	Resultados observados	20
5	Conclusões	23
5.1	Trabalhos futuros	25
Referências		27

Capítulo 1

Introdução

Nos últimos anos, vivemos um momento de crescente participação popular na política brasileira. Com a crescente demanda por transparência na gestão pública, os governos assumem o compromisso de tornar públicos, através da internet, os dados de suas gestões - como arrecadações, gastos, licitações e dados de serviços públicos. Surge então um meio capaz de diminuir as barreiras para que a população fiscalize a atuação dos nossos representantes.

Entre os compromissos da transparência, está a disponibilização de dados abertos governamentais (DAGs), que consiste na distribuição de qualquer tipo de dado em posse do poder público em formatos que permitem qualquer cidadão consultar, analisar, se apropriar e redistribuí-los de maneira livre. Esses dados permitem não só uma maior fiscalização da administração pública, como também permitem que indivíduos e organizações possam utilizá-los para construir ferramentas de valor à sociedade.

No município do Rio de Janeiro, entre os diversos dados abertos oferecidos, estão os dados de transporte e mobilidade, que incluem informações sobre linhas de ônibus, metrô, trem, bicicletários e ocorrências de trânsito. Neste trabalho, serão utilizados os dados disponibilizados das linhas de ônibus municipais, que incluem informações dos GPS dos ônibus, a fim de observar o comportamento desses veículos e analisar o funcionamento desse serviço público na cidade.

1.1 Objetivos

O objetivo deste trabalho é elaborar uma ferramenta para o cidadão carioca poder fiscalizar a qualidade de serviço do transporte público rodoviário da sua cidade, através da análise dos dados reais obtidos pelos DAGs locais. Tal ferramenta permite ao usuário acessar relatórios diversos para auxiliar na compreensão desse serviço público prestado. Este trabalho foca no desenvolvimento de duas análises específicas: a densidade de ônibus por região e a frequência média dos ônibus de cada linha.

A primeira análise foi selecionada devido à uma reflexão dos cariocas: "Por que há lugares que possuem tantos ônibus e outros que não são atendidos por nenhuma linha?". O estudo proposto serve para expor esse grande contraste e desigualdade presentes na cidade.

Já a segunda análise foi inspirada por uma das maiores dúvidas de qualquer usuário de ônibus: "Quanto tempo vai demorar para o meu ônibus chegar?". Para ajudar a responder esta pergunta, calcularemos os intervalos entre cada ônibus ao longo de um período, comprovando qual é o tempo médio de espera por cada linha.

1.2 Motivação

A ideia para este trabalho surgiu durante minha participação no Rio Bus[1] - um aplicativo móvel desenvolvido na Universidade Federal do Rio de Janeiro que utiliza DAGs para informar a localização dos ônibus de cada linha do Rio de Janeiro, servindo como grande utilitário para o usuário de ônibus carioca.

Esses usuários, muitas vezes por engano, enviavam para nós reclamações sobre o serviço dos ônibus - algo sobre o qual não temos responsabilidade. No entanto, pensamos em uma maneira de ajudar: fornecendo ao usuário relatórios de funcionamento dos ônibus e munindo-o com informações concretas para que possa enviar suas reclamações aos órgãos fiscalizadores.

Em paralelo, também houve uma procura pelos dados dos ônibus armazenados

no Rio Bus por políticos que promoviam a CPI dos Ônibus[2]. Estes buscavam uma fonte alternativa aos burocráticos processos legais que fosse capaz de fornecer relatórios como evidência do descumprimento dos contratos por parte das concessionárias que operam as linhas de ônibus. Este foi um dos diversos pontos questionados na CPI, buscando apurar a eficácia dos instrumentos de controle das metas estabelecidas nos contratos com essas empresas.

Assim, graças ao transporte público extremamente deficiente do Rio de Janeiro e essa demanda por informação - que, idealmente, deveria ser pública - decidi desenvolver este projeto para tentar contribuir, de alguma maneira, com a melhoria desse essencial serviço.

Ainda que não esteja em minhas mãos garantir que os usuários irão fazer uso desses dados para reivindicar seus direitos - e muito menos garantir que essas melhorias serão feitas - gostaria de contribuir com ferramentas que possam auxiliar isso. Além disso, todos os estudos e algoritmos desenvolvidos durante este projeto são *open-source*, ou seja, de domínio público¹.

1.3 Trabalhos relacionados

Trabalhos anteriores buscaram analisar a definição e os impactos dos dados abertos na sociedade. Em [4], o autor descreve os fundamentos dos DAGs em quatro interpretações - como um direito, como movimento e como política - e analisa os impactos causados na *accountability*, na melhoria dos serviços públicos e no crescimento econômico. Em [5], o autor traz um estudo de caso do Buus, uma empresa do Rio de Janeiro que faz uso dos mesmos DAGs que utilizaremos e que acompanhou o processo de abertura dos dados pela Prefeitura do Rio. São estudadas na prática as dificuldades e consequências dos usos desses dados, assim como diferentes modelos de negócios testados durante o desenvolvimento.

¹"Código aberto, ou open source em inglês, é um modelo de desenvolvimento que promove um licenciamento livre para o design ou esquematização de um produto, e a redistribuição universal desse design ou esquema, dando a possibilidade para que qualquer um consulte, examine ou modifique o produto", de acordo com [3].

O próprio Google também possui ferramentas que fazem uso dos DAGs do Rio de Janeiro. Sua principal ferramenta de navegação, o Google Maps, oferece informações sobre o transporte público do Rio e estimativas do tempo de chegada de ônibus próximos². Entretanto, esse serviço difere do desenvolvido neste trabalho pois o mesmo utiliza apenas uma estimativa futura para os ônibus mais próximos durante a pesquisa, enquanto este projeto faz a análise histórica dos dados e, portanto, efetiva dos ônibus. Além disso, este projeto torna possível a análise retroativa sobre qualquer período anterior, enquanto na ferramenta norte-americana só é possível consultar para o instante de tempo atual.

1.4 Estrutura e metodologia

No capítulo 2, são apresentados alguns dos conceitos fundamentais que cercam o projeto, como Dados Abertos e seu contexto no cenário em que estamos interessados - os Dados Abertos Governamentais. No capítulo 3, apresentaremos o cenário do Rio Bus, projeto que serviu como base para as pesquisas desenvolvidas, explicando como foram obtidos, armazenados e manipulados os dados que utilizamos. O capítulo 4 apresenta as análises sobre os dados, incluindo suas implementações e os resultados observados. O capítulo 5 contém as conclusões finais sobre o projeto e trabalhos futuros.

²"Ônibus do Rio de Janeiro aparecem no Google Maps em tempo real", TechTudo, 11/04/2016, <http://www.techtudo.com.br/noticias/noticia/2016/04/onibus-do-rio-de-janeiro-aparecem-no-google-maps-em-tempo-real.html>

Capítulo 2

Conceitos

Dados Abertos

"Dado abertos", segundo a *Open Definition* (OD) da *Open Knowledge Foundation* [6], são dados disponibilizados por organizações, empresas e indivíduos cujo conteúdo pode ser livremente acessado, utilizado, modificado e redistribuído por qualquer um e com qualquer propósito, estando sujeito a, no máximo, a exigência de creditar sua autoria e compartilhar pela mesma licença.

Ainda segundo a OD, há caracterizações específicas a respeito do acesso, reutilização, tecnologia e outros aspectos que garantem a abertura dos dados. Quanto ao acesso e tecnologia, os dados devem ser disponibilizados sempre na íntegra e, no máximo, a um custo razoável de reprodução, sendo preferencialmente distribuído de forma gratuita pela internet. Devem estar disponíveis de forma conveniente e modificável em formato aberto, cuja especificação é pública e não impõe restrições monetárias e outras à sua utilização (como, por exemplo, formatos de softwares proprietários).

Sobre o uso, a licença deve permitir modificações e trabalhos derivados e deve ser permitida a distribuição sob os termos do trabalho original. Ela não deve restringir ninguém de redistribuir - gratuitamente ou não - o conteúdo, seja ele sozinho ou como parte de um coletivo compilado a partir de diferentes fontes. Esta regra que permite o desenvolvimento de estudos e ferramentas - como o desenvolvido neste

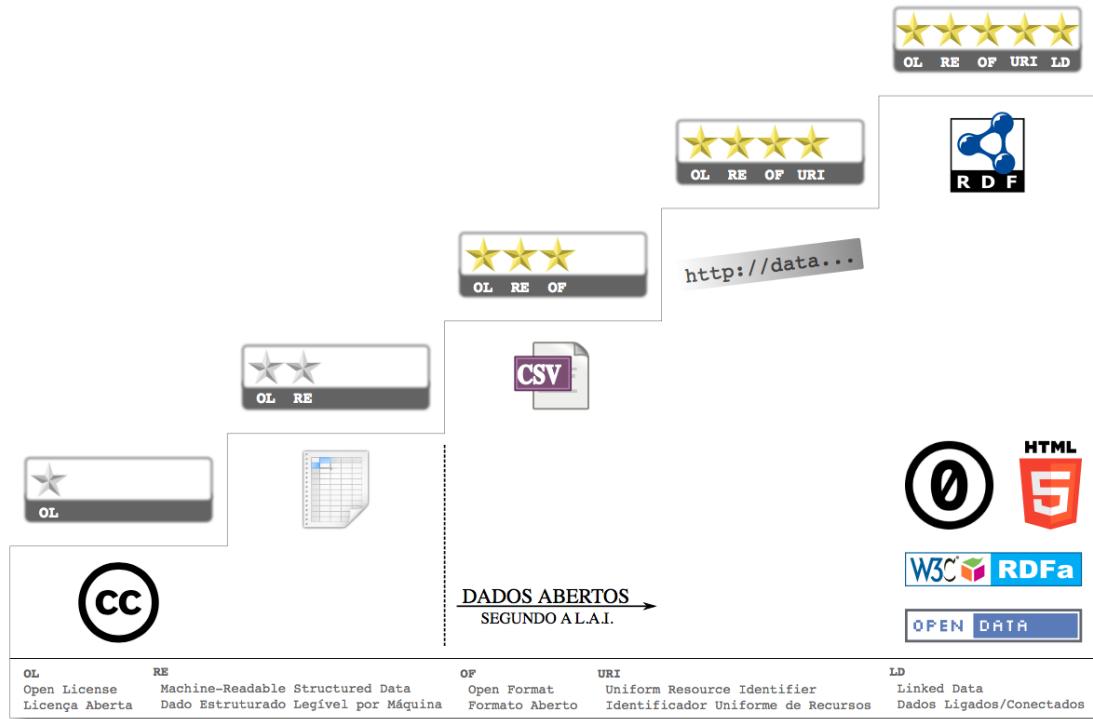


Figura 2.1: Classificação dos dados abertos segundo a escala de 5 estrelas de Berners-Lee

trabalho - de forma gratuita, sem o custo de royalties e taxas sobre a redistribuição da informação utilizada.

Como nem todos os bancos de dados abertos atendem a todos os requisitos elaborados, foi definida uma escala de 1 a 5 estrelas para diagnosticar a abertura dos dados de acordo com quantos princípios da OD são implementados[7]. Essa escala, criada por Tim Berners-Lee, criador da *World Wide Web*, pode ser vista na Figura 2.1. De acordo com a proposta, poderíamos diagnosticar um dataset nos seguintes níveis:

- **1 estrela:** Disponibilizar o dado na web (em qualquer formato) sob uma licença aberta.
- **2 estrelas:** Tornar o dado estruturado (exemplo: uma planilha no Excel ao invés de uma imagem digitalizada de uma tabela).
- **3 estrelas:** Tornar o dado disponível em um formato aberto, não-proprietário (exemplo: CSV ao invés de XLS, do Excel).

- **4 estrelas:** Utilizar URIs para denotar os dados, para que as pessoas possam referenciá-los.
- **5 estrelas:** Conectar os dados a outros datasets a fim de fornecer um contexto (*linked data*¹).

Dados Abertos Governamentais

Apesar de as definições de dados abertos serem gerais o suficiente para qualquer contexto, foram elaborados princípios mais específicos como referência para dados abertos governamentais, isto é, dados abertos sob posse de um governo[8]. Um grupo de trabalho de 30 pessoas situado na Califórnia, Estados Unidos, o *Open Government Working Group*, chegou a um consenso sobre os seguintes princípios:

1. **Completos.** Todos os dados públicos são disponibilizados. Dados são informações eletronicamente gravadas, incluindo, mas não se limitando a, documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, reguladas por estatutos.
2. **Primários.** Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível, e não de forma agregada ou transformada.
3. **Atuais.** Os dados são disponibilizados o quanto rapidamente seja necessário para preservar o seu valor.
4. **Acessíveis.** Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis.
5. **Processáveis por máquina.** Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado.

¹O conceito de linked data (dados ligados entre si) é um conjunto de práticas introduzidas por Tim Berners-Lee, com função de publicar e estruturar dados na Web, agregando valor semântico aos dados.

6. **Acesso não discriminatório.** Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro.
7. **Formatos não proprietários.** Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo.
8. **Livres de licenças.** Os dados não estão sujeitos a regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos.

Uma vez tendo em mãos esses dados, há diversos possíveis usos dos mesmos refletindo em diferentes aspectos da sociedade. Os DAGs trazem impacto direto na política e são causadores de um grande movimento de *accountability*, isto é, uma responsabilidade aos governos a partir das informações oriundas desses dados. Dessa forma, podemos enxergar que quanto maior o acesso a essas informações, maiores as chances de que a sociedade acompanhe as iniciativas do poder público e as fiscalize, eventualmente contribuindo para a melhora do serviço público.

Outro grande reflexo dos DAGs é em termos de mercado. A partir da mineração dos dados abertos, que vêm como dados brutos, é possível extrair novas informações e agregá-las na forma de produtos e serviços de valor para os usuários. Tem-se então a criação de novos negócios a partir dos DAGs. Alguns dos possíveis modelos de negócios a partir de DAGs e seus impactos na economia foram estudados por [5].

Capítulo 3

Rio Bus

Uma aplicação real que faz uso dos DAGs e que serviu como plataforma para este trabalho é o **Rio Bus**[1], projeto do qual faço parte. Criado por alunos do curso de Bacharelado em Ciência da Computação da UFRJ, o Rio Bus surgiu em 2014 como um aplicativo para visualizar em um mapa a posição em tempo real dos ônibus municipais do Rio de Janeiro, utilizando os recém-lançados dados abertos de mobilidade oferecidos pela Prefeitura do Rio de Janeiro. Com esse serviço, o cidadão pode estimar o quanto vai demorar para seu ônibus chegar e, assim, se planejar melhor.

Lançado gratuitamente e como um projeto de *open-source*, o Rio Bus logo conseguiu vários colaboradores, usuários do aplicativo, dentro da UFRJ, haja vista a grande demanda por transporte público no campus da Universidade. A solução foi então disponibilizada em diferentes plataformas - um *website*, disponível em [9], para acesso pelo computador ou dispositivos móveis, e dois aplicativos para *smartphones*,



um para o sistema operacional *Android*, do Google, disponível em [10], e outro para *iOS*, da Apple, disponível em [11].

Pouco tempo após seu lançamento, o aplicativo ganhou uma grande repercussão e serviu como ferramenta para muitos cidadãos durante a greve dos ônibus, em março de 2014. Durante a paralisação, na qual cerca de 80% da frota de ônibus do município não saiu da garagem das empresas, muitos usuários fizeram uso do aplicativo para saber se suas linhas de ônibus estavam circulando e acompanhar onde estavam os coletivos. Esse evento não só serviu como a tração que o Rio Bus precisava para seu sucesso, como também foi um grande *insight* de que coletar os dados que vinham sobre os ônibus poderia servir para uma análise do funcionamento do serviço. Esses estudos não só ajudaram a melhor compreender o serviço dos ônibus no Rio e suas falhas, como também podem ter importância especial durante períodos atípicos como paralisações, alterações no sistema, grandes eventos e ocorrências de trânsito na cidade.

3.1 Base de dados

3.1.1 Origem

Para popular a base de dados do Rio Bus, são utilizados os dados dos ônibus fornecidos pelo Data.rio[12], o portal de dados abertos da Prefeitura do Rio de Janeiro. O Data.rio disponibiliza um total de 1200¹ conjuntos de dados (*datasets*) que estão classificados em 13 grupos - entre eles, Urbanismo, Administração Pública, Impostos e Transporte e Mobilidade, que foi o grupo utilizado.

Dentre os dados de Transporte e Mobilidade, há 23 conjuntos de dados que incluem informações sobre ônibus, trens, metrô, barcas e bicicletários do Rio de Janeiro. Estamos interessados em 3 desses *datasets*: GPS dos ônibus, Pontos de parada das linhas do ônibus e Pontos dos trajetos das linhas de ônibus. Esses dados são fornecidos em parceria da Prefeitura do Rio com a FETRANSPO (Federação das Empresas de Transporte de Passageiros do Estado do Rio de Janeiro).

¹Número de conjuntos de dados disponíveis em <http://data.rio/> em 6 de julho de 2016

Tabela 3.1: Exemplo da coleção histórica de um ônibus

Trecho da coleção histórica do ônibus B28518 da linha 325						
Data/hora	Ordem	Linha	Lat.	Long.	Veloc .	Direção
09/07/2015 23:19:05	B28518	325	-22.815200	-43.188300	41	306
09/07/2015 23:20:18	B28518	325	-22.815700	-43.188100	20	165
09/07/2015 23:21:35	B28518	325	-22.815700	-43.188000	37	150

3.1.2 Formatos

São utilizados diferentes formatos para acesso aos dados do Data.rio. Os dados dos pontos de parada e pontos de trajeto das linhas dos ônibus são dados tipicamente estáticos - isto é, não costumam sofrer alterações frequentes - e são disponibilizados em uma planilha em formato CSV (Comma-separated values). Eles são atualizados com baixa frequência na base de dados do Rio Bus.

Já os dados de GPS dos ônibus, como são atualizados diversas vezes por minuto, podem ser acessados através de uma API (Application Programming Interface) que retorna um objeto JSON (JavaScript Object Notation) contendo os dados de cada ônibus naquele instante. Esses dados são coletados com alta frequência para popular a base de dados do Rio Bus, formando assim uma coleção histórica dos dados. Um exemplo de como se parece essa coleção está na tabela 3.1, que mostra um trecho dos registros de um ônibus da linha 325 no dia 09/07/2015.

3.1.3 Armazenamento

Como o volume de dados proveniente da atualização dos ônibus é muito grande, as APIs do Data.rio fornecem apenas os dados do instante atual. Para que pudessemos ter acesso aos dados de períodos passados, utilizamos a infraestrutura do *back-end* do Rio Bus para armazenar esses dados, que ultrapassam 9 milhões de

novos registros por dia.

Para isso, temos um serviço que executa indefinidamente que faz a leitura dos dados do Data.rio e os salva num banco de dados próprio. Na implementação atual, os dados são salvos em dois lugares: numa coleção do MongoDB, um banco de dados não-relacional open-source instalado localmente, e no Google BigQuery, um serviço de cloud para armazenamento e análise de dados em larga escala (*Big Data*), como é o nosso caso.

Para os propósitos deste trabalho, daremos preferência ao banco de dados MongoDB[13] para executar as consultas. Ainda que a implementação das consultas nele seja mais difícil - o Google BigQuery[14] suporta consultas com SQL - a decisão leva em conta o acesso à informação - o BigQuery é um serviço pago, remoto e de código fechado, enquanto o MongoDB é uma aplicação que pode ser executada localmente, sem custos e que possui o código aberto, de acordo com os princípios deste projeto.

Capítulo 4

A análise dos dados

O escopo inicial deste trabalho propôs 4 diferentes análises:

1. **Frequência da linha:** calcula qual foi a frequência e tempo médio de espera por ônibus de uma linha. Reflete quanto tempo um usuário costuma esperar, em média, até que passe seu ônibus, em cada horário do dia. Pode ser usada para comparação com a frequência informada por cada companhia e para o próprio usuário poder estimar o tempo que irá aguardar até embarcar.
2. **Concentração da frota:** mostra onde os ônibus estiveram nas últimas 24 horas, mostrando quais regiões possuem maior concentração de ônibus e quais sequer possuem linhas de ônibus.
3. **Quantidade de ônibus por linha e hora:** similar ao estudo anterior, porém mostra quantos carros de cada linha estavam circulando em cada horário do dia. Pode ser comparada com a quantidade de veículos determinada nas concessões da Prefeitura às companhias de ônibus.
4. **Atualização do GPS:** uma meta-análise dos DAGs para monitorar a qualidade do monitoramento dos ônibus em relação à atualização dos dados. Pode servir pra mostrar quais companhias não estão transmitindo os dados de sua frota e, portanto, descumprindo seu contrato com a Prefeitura.

Foram desenvolvidos os estudos 1 e 2, que serão descritos em detalhes a seguir.

4.1 Frequência dos ônibus

Este estudo tem como objetivo analisar a frequência dos ônibus de uma linha, de maneira que o resultado final reflita qual o tempo médio de espera por um ônibus daquela linha em um certo horário do dia. Tal informação não é oferecida pelos DAGs utilizados e, apesar de algumas companhias de ônibus divulgarem uma frequência oficial de suas linhas, muitas vezes estas não são cumpridas ou, na prática, são afetadas por outros fatores, como o trânsito.

4.1.1 Dados utilizados

As estatísticas são obtidas a partir de uma amostra dos registros de localização dos ônibus em um período de 24 horas. Como as linhas possuem diferentes características, cada linha foi analisada individualmente e os dados médios serão agrupados por hora.

Também são usados para compreensão desses dados a informação geolocalizada dos pontos de parada da linha.

4.1.2 O algoritmo

Calcularemos o intervalo de tempo entre cada ônibus da linha em diversos pontos ao longo do itinerário. Considere disponíveis as variáveis e funções abaixo:

- L : a linha a ser analisada (ex.: 485);
- D : a data a ser analisada (ex: 08/10/2015);
- $stops(L)$: função que retorna o conjunto de pontos de parada geolocalizados da linha L ;
- $history(L, D)$ a função que retorna os registros geolocalizados dos ônibus da linha L na data D ;
- $matches(R, p)$ a função que retorna o subconjunto dos registros R próximos a um ponto de parada p .

O algoritmo desenvolvido é descrito a seguir.

1. Obter $P = stops(L)$ e $R = history(L, D)$.
2. Para cada ponto p_i de P :
 - 2.1. Obter $M_i = matches(R, p_i)$, isto é, os registros dos ônibus que passaram pelo ponto de parada p_i .
 - 2.2. Ordenar M_i pelo horário dos registros de forma ascendente.¹
 - 2.3. Eliminar de M_i registros de um mesmo ônibus com horários próximos, a fim de evitar que ele seja contado mais de uma vez.²
 - 2.4. Dividir os cruzamentos de M_i em 24 subconjuntos M_{ij} de acordo com a hora de cada registro ao longo do dia.
 - 2.5. Para cada conjunto M_{ij} , $0 \leq j < 24$, calcular o intervalo de tempo entre cada registro do conjunto e a média I_{ij} dos intervalos.
3. Para cada hora do dia, $0 \leq j < 24$, calcular a média entre todos os pontos i dos intervalos I_{ij} .

Ao fim do algoritmo, obtém-se um conjunto com 24 médias, representando o intervalo médio - ou seja, a frequência - dos ônibus daquela linha em cada hora do dia.

O motivo de a frequência dos ônibus ser analisada por hora é devido ao próprio modelo de funcionamento das linhas, que possui frequências variáveis de acordo com a demanda de cada horário.

Através do algoritmo mostrado, também é possível calcular qual foi o tempo de retorno do ônibus, indicando o tempo total de viagem na linha (ida + volta).

¹Obtém-se ao final da ordenação a sequência de cada ônibus que passou por aquele ponto ao longo do dia - assim como teria, por exemplo, um fiscal da companhia de ônibus, cuja função é anotar os horários em que cada ônibus da companhia passou por ali.

²Caso o ônibus encontre-se parado durante muito tempo no ponto ou próximo a ele, pode haver mais de um registro, cujos cruzamentos duplicados devem ser descartados.

Para isso, basta calcular o tempo entre duas ocorrências do mesmo ônibus no mesmo ponto e com o mesmo sentido - ou seja, o tempo que o ônibus levou desde que passou ali até percorrer a linha toda e passar ali de novo.

4.1.3 Implementação

O algoritmo foi implementado em Node.js, que é um ambiente de desenvolvimento em JavaScript[15][16]. Essa escolha foi feita principalmente devido à velocidade do desenvolvimento e por ser uma boa linguagem de prototipagem[17] e possui um excelente suporte ao MongoDB, banco de dados utilizado.

Foi desenvolvido um programa que faz a conexão com o banco de dados, já populado, executando o algoritmo acima e exportando os dados calculados para um arquivo JSON.

Também foi criada uma página HTML[18] que usa JavaScript para ler os dados do arquivo JSON e plotá-los num gráfico de barras, a fim de facilitar a visualização dos dados de hora em hora ao longo do dia analisado.

4.1.4 Resultados observados

Foi utilizado como exemplo do estudo o histórico de julho de 2015 da linha 324, que faz o percurso Ribeira (Ilha do Governador) x Candelária (Centro). Segundo o site da empresa que controla essa linha, a Viação Ideal, a frequência média oficial é de 9 minutos entre cada ônibus[19].

Na figura 4.1, os dados analisados são de um dia útil. Podemos observar durante a manhã desse dia que o tempo médio de espera por um ônibus dessa linha fica em torno de 20 minutos, com pequenos aumentos chegando a 25 minutos para quem aguardava um ônibus entre as 8 e 10 horas, período com reflexos do horário do rush. Já na faixa das 12 horas, nenhuma média foi calculada, pois não houve frequência suficiente de ônibus nos pontos analisados para que o algoritmo pudesse calcular uma média. Durante o começo da tarde, vemos um grande aumento no tempo de espera, provavelmente causado pelo mesmo motivo que causou a falta de dados no

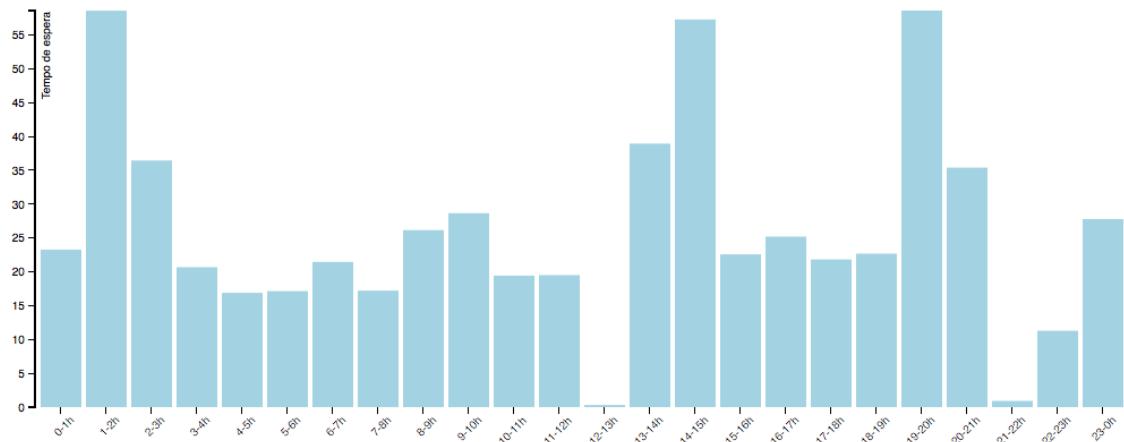


Figura 4.1: Análise mostra o tempo médio de espera por um ônibus da linha 324 num dia útil em julho de 2015.

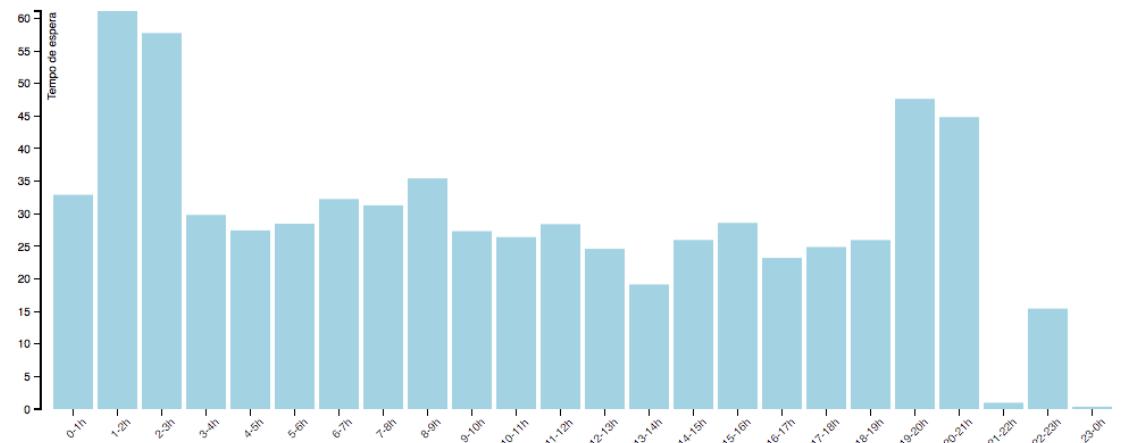


Figura 4.2: Análise mostra o tempo médio de espera por um ônibus da linha 324 num domingo em julho de 2015.

período das 12 horas, que foi normalizando nas horas seguintes até que aumentou de novo a partir das 19 horas.

Uma comparação da análise desse dia pode ser feita com a da figura 4.2, com os dados da mesma linha porém em um domingo. Duas observações podem ser feitas imediatamente: a média de espera é maior, pois aos finais de semana há menos ônibus circulando e também o tempo de espera é bem mais constante, visto que num domingo não há trânsito causando atrasos na circulação dos ônibus.

Com uma base de dados maior, é possível fazer tais comparações entre datas mais relevantes, a fim de observar os reflexos na frequência dos ônibus causados por eventos, interdições, desvios e outros fatores. Também é possível agregar dados analisados em intervalos maiores.

Apesar de o algoritmo ser relativamente simples, houve uma grande dificuldade na filtragem dos registros duplicados. Ao observar que os resultados da execução com uma certa linha estavam apontando um resultado surpreendentemente bom, notamos que existe um cenário específico em que os ônibus são erroneamente identificados cruzando um ponto de parada.

Foi comparada a execução do algoritmo em diferentes linhas com resultados similares. Concluiu-se que estava sendo computado um falso positivo dos ônibus que cruzavam um ponto de parada no sentido contrário ao esperado. Logo, observou-se que isso era um cenário recorrente nas linhas em que o itinerário de ida era similar ao itinerário de volta, como quando uma linha passa na mesma rua nos dois sentidos. Nesse caso, os pontos de parada da ida são muito próximos aos pontos da volta - às vezes com pouquíssimos metros de distância, apenas em lados opostos da rua.

O algoritmo que calcula o cruzamento de um ônibus com um ponto de parada através de seu raio de proximidade acabava identificando um falso positivo nesses casos. Visando contornar esse problema, foi necessário implementar essa diferenciação do sentido por conta própria.

A fim de identificar se um ônibus está no seu itinerário de ida ou de volta, foi elaborado um algoritmo que identifica qual o sentido comparando a posição atual do ônibus com a sua posição nos últimos registros. Considerando que conhecemos o itinerário da linha, comparamos o histórico das últimas n posições conhecidas do ônibus, guardadas na memória, com o itinerário dos dois sentidos, analisando se o trajeto pertence ao conjunto da ida ou da volta.

Uma vez conhecido o sentido atual, esse dado é salvo junto ao registro do ônibus para que possa ser utilizado posteriormente. Para identificar se um determinado ônibus realmente passou por um ponto de parada, verificamos se o sentido registrado naquele instante condiz com o sentido do ponto. Caso sejam diferentes, esse registro é descartado, eliminando o falso positivo.

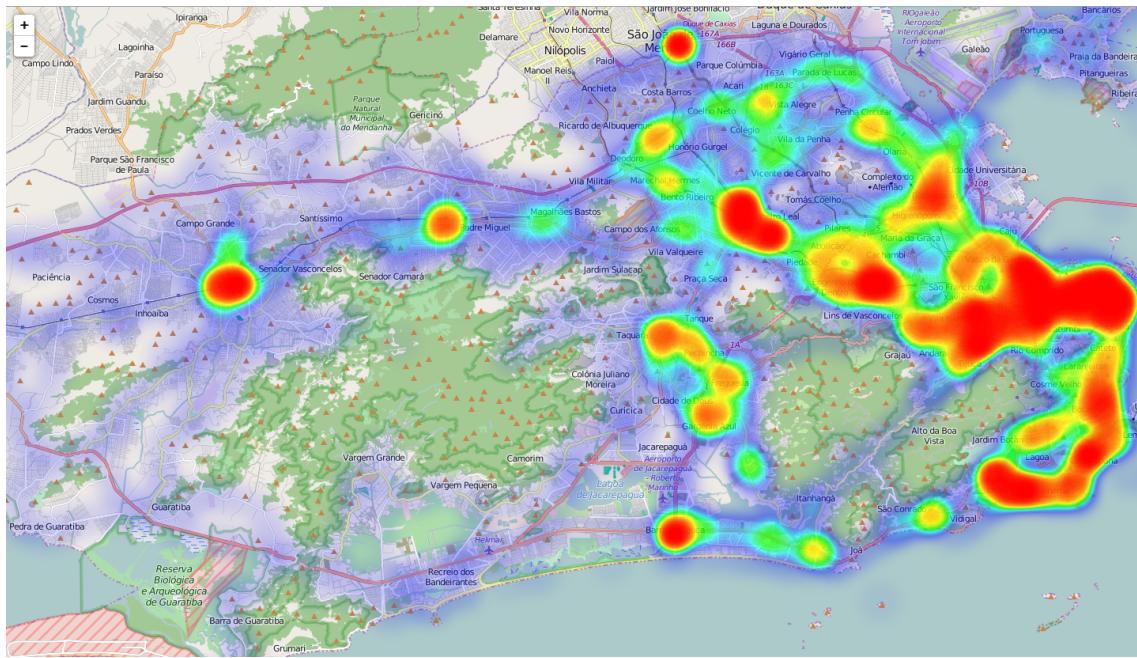


Figura 4.3: Vista geral da cidade mostra uma grande concentração de ônibus nas regiões centrais. (Outubro de 2015)

4.2 Concentração da frota

A proposta deste estudo é evidenciar, através de um mapa de calor, onde estão concentrados os ônibus pela cidade ao longo de um dia. Pode ser usada como ferramenta para o planejamento de novas linhas e redistribuição das atuais.

Apesar de ser difícil extrair conclusões apenas com este mapa, ele pode servir como uma ferramenta visual para observar em que vias e regiões da cidade estão concentrados os ônibus. Um exemplo é a figura 4.3, que traz uma visão geral da cidade do Rio de Janeiro.

4.2.1 Implementação

Por ser uma análise mais simples, o esforço maior desta análise foi na parte da apresentação dos dados, ao invés da extração dos mesmos.

Para extrair os dados, foi feita uma consulta no banco de dados por todos os registros geolocalizados em uma data específica. Como extraír e plotar esses dados diretamente seria muito demorado - há mais de 15 milhões de registros por dia

- optamos por reduzir manualmente a precisão desses dados, agrupando registros muito próximos.

Para agrupar essas amostras, limitamos na nossa consulta a precisão dos campos de latitude e longitude de cada registro, utilizando uma precisão máxima de 4 casas decimais ao invés de 6. Só isso já reduz nosso número de amostras em cerca de 35 vezes, para cerca de 420.000 amostras, ainda mantendo uma boa precisão visual dos dados.

Finalmente, a fim de descobrir quantos ônibus estavam em cada coordenada, foi utilizada a quantidade de registros naquela posição, contando apenas veículos distintos (cada ônibus é identificado por um número de ordem). Temos então como saída uma lista de coordenadas e suas respectivas quantidades de ônibus.

Cada coordenada é plotada no gráfico através de uma escala de cores que representa a densidade de ônibus naquela região[20]. Como resultado, temos cores entre o azul escuro, que representa uma baixa densidade de ônibus na região, e o vermelho, que representa uma alta densidade. As escalas de cores foram normalizadas entre os estudos para preservar facilitar a comparação entre diferentes mapas.

Para a visualização desses dados, foi elaborada uma página que renderiza um mapa interativo, no qual é possível alterar a região de estudo e observar com maior precisão os resultados, permitindo focar em bairros ou ruas específicas. Quando duas amostras são comparadas, é possível visualizá-las lado-a-lado, atualizando simultaneamente conforme o mapa é movimentado pelo usuário.

4.2.2 Resultados observados

O mapa de calor se torna ainda mais útil quando são comparados dados de diferentes períodos, a fim de analisar como mudou a distribuição dos ônibus ao longo do tempo.

Uma das comparações feitas foi entre dados de dois anos consecutivos, 2015 e 2016. Foram analisados dados do mesmo dia, 8 de julho, um dia útil, nos dois anos. A comparação entre esses anos é especialmente relevante pois, desde o final de 2015

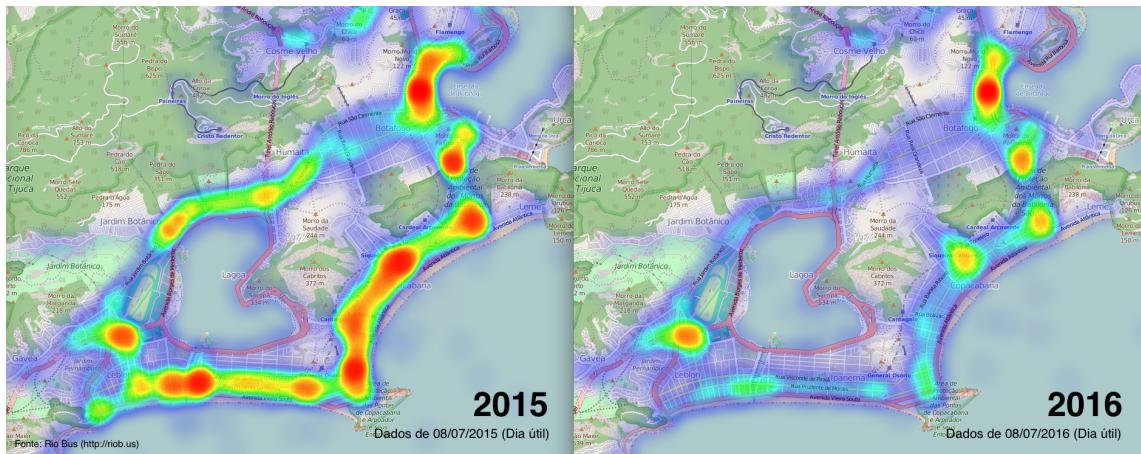


Figura 4.4: Comparação da quantidade de ônibus na Zona Sul em dois anos consecutivos. Em 2016, observa-se a redução em virtude da racionalização das linhas de ônibus.

até a data atual, ocorrem grandes mudanças nas linhas de ônibus do Rio de Janeiro, em especial devido ao plano de racionalização das linhas da Zona Sul[21]. Além disso, devido a grandes obras pela cidade, houve diversas interdições no trajeto das linhas, a construção de novas vias e implantação de novos corredores exclusivos de ônibus.

Na figura 4.4, é possível visualizar essa comparação e observar uma enorme diminuição dos ônibus nos bairros da Zona Sul próximos à Lagoa Rodrigo de Freitas.

No Centro da cidade, observa-se através da figura 4.5 que as principais vias da região concentram praticamente toda a circulação dos ônibus, enquanto vias transversais não possuem quase nenhum tráfego desse tipo.



Figura 4.5: No Centro, concentração alta nas vias principais e baixa nas transversais. (Outubro de 2015)

Capítulo 5

Conclusões

Com os estudos realizados, foi possível construir mais uma prova de como os DAGs podem ser úteis à sociedade. As duas análises desenvolvidas são exemplos de como esse tipo de ferramenta pode ser usada para ampliar a transparência de serviços como o transporte público. Porém, ao mesmo tempo que temos que agradecer pela disponibilidade desses dados públicos, ainda há muito o que se reivindicar no que diz respeito à qualidade e à cobertura dos dados fornecidos, visto que existem diversos problemas que foram observados durante os estudos.

Foram identificadas algumas dificuldades em relação à confiabilidade dos dados coletados, uma vez que a base de dados que utilizamos da Prefeitura não possui informações completas sobre as linhas. Além disso, também há uma enorme quantidade de dados desatualizados, o que impede a execução correta dos algoritmos para a análise, causando a disponibilização de informações incorretas à população que utiliza serviços como o Rio Bus, por exemplo.

Há alguns exemplos de como tais problemas afetaram este estudo:

- Diversas linhas em circulação não possuem informações sobre seu itinerário, um dos dados oferecidos pelo Data.rio, enquanto diversas outras possuem itinerários desatualizados. Sem esses dados não há uma referência para a comparação das linhas através de algoritmos.

- Assim como o itinerário, faltam informações sobre os pontos de parada ou estão desatualizadas em diversas linhas, o que impede a execução da análise da frequência dos ônibus ou causa resultados não confiáveis.
- Mudanças que não existiam ou que ocorriam em menor escala nas linhas de ônibus no Rio de Janeiro ocorreram com grande frequência no último ano, dificultando os estudos, tais como a mudança de 180 linhas que passam pelo Centro do Rio em maio de 2016[22] e o plano de racionalização das linhas, que reduziu de 123 para 45 as linhas que passam pela Zona Sul do Rio[21].
- Não é informado pelos DAGs qual o sentido do ônibus em cada instante, o que acaba tendo de ser calculado manualmente para que essa informação seja usada para calcular com maior precisão o tempo de espera em cada parada.
- Há falhas frequentes na transmissão dos dados de posição dos ônibus, com linhas ou até mesmo consórcios inteiros¹ que "desaparecem" durante períodos, conforme observado em 2015, quando o consórcio Internorte ficou meses sem transmitir dados de todas as suas linhas.
- A própria API do Data.rio possui problemas, ocasionalmente ficando indisponível sem nenhum aviso prévio e alterando o contrato de seus serviços sem a documentação apropriada, o que causou interrupções e falhas na coleta dos dados.

Devido aos problemas relatados anteriormente, a complexidade do trabalho se tornou maior, pois foi necessário elaborar soluções além do escopo inicial para que se obtivesse maior confiabilidade dos resultados, causando interrupções no desenvolvimento dos estudos e criando limitações nas linhas que poderiam ser analisadas.

Outros problemas encontrados foram em relação ao armazenamento dos dados e à complexidade de processamento dos mesmos. Com a enorme quantidade de

¹As empresas de ônibus que operam no Rio de Janeiro são agrupadas em 4 diferentes consórcios - Internorte, Intersul, Transcarioca e Santa Cruz - que operam em diferentes áreas da cidade. Fonte: Rio Ônibus (<http://www.rioonibus.com/rio-onibus/consorcios-e-empresas/>). Acessado em outubro de 2016.

dados produzida a cada dia, armazenar dados históricos localmente tornou-se um problema em diversos aspectos, como o uso de espaço em disco, a disponibilidade de servidores capazes de colher esses dados ininterruptamente e a complexidade de processamento dos dados. Embora a infraestrutura do Rio Bus e seus serviços, de código aberto, tenham ajudado nesse aspecto, ainda dependíamos de serviços externos como o Google BigQuery para consultar esses dados históricos, o que gerou custos e mais um nível de dependência.

Quanto à complexidade de processamento, processar dados de longos períodos ou de muitas linhas simultaneamente foi bastante demorado. Por este motivo, o estudo da frequência dos ônibus no escopo deste trabalho limitou-se a apenas uma linha e a um dia por vez, embora pudesse ser otimizado para análises mais elaboradas.

Apesar de todos os desafios encontrados, a ferramenta desenvolvida serviu como uma prova de conceito de que é possível utilizar os DAGs como instrumento para a fiscalização do serviço de transporte no Rio. Com as análises desenvolvidas, foi possível observar que o serviço prestado deixa muito a desejar, e como as mudanças nos planos de transporte da cidade afetam a disponibilidade das linhas.

Ainda que o objetivo inicial deste trabalho propusesse uma facilidade maior de acesso às estatísticas coletadas, todos os algoritmos desenvolvidos durante os estudos foram disponibilizados no GitHub², podendo ser executados, modificados ou redistribuídos por qualquer um, assim como o código do Rio Bus, que serviu como uma útil ferramenta para a coleta dos dados abertos.

5.1 Trabalhos futuros

Apesar de o projeto ter sofrido limitações devido à base de dados utilizada, suas propostas e algoritmos podem ser facilmente reproduzidos em outras bases de dados, servindo a diferentes propósitos e podendo ser facilmente estendidos e aprimorados

²GitHub: site que permite que as pessoas hospedem e compartilhem códigos versionados, e concentra o maior número de projetos open-source disponíveis, com uma enorme comunidade de contribuidores. (<https://github.com>)

para outros meios de transporte.

Substituindo a base de dados utilizada, é possível reproduzir os estudos em diferentes cidades que disponibilizam os dados dos ônibus como DAGs e até mesmo em diferentes meios de transportes. Embora o estudo tenha sido pensado com o funcionamento dos ônibus públicos, as mesmas análises podem ser aplicadas a frotas de ônibus privados, táxis, carros ou caminhões. Um serviço de monitoramento e de relatórios gerenciais de frota pode ser aplicado para fins comerciais e oferecido a cooperativas, transportadoras e empresas de ônibus, como é apontado por [5].

Com a infraestrutura adequada, o cálculo das análises também pode ser automatizado para que, por exemplo, a frequência de cada linha seja calculada ao final de cada dia. Dessa maneira, os estudos desenvolvidos podem se tornar ainda mais completos, e os resultados podem ser disponibilizados em uma página abrangendo todas as linhas de ônibus.

O algoritmo de análise da frequência também pode ser explorado com outras abordagens de implementação, como, por exemplo, aprimorando o paralelismo do algoritmo. Tal mudança melhoraria o desempenho de processamento, algo especialmente importante caso se queira analisar o conjunto de todas as linhas de maneira eficiente.

Ainda visando uma ferramenta pública para a consulta desses dados, pode ser criado um repositório contendo as diferentes análises propostas, para que qualquer cidadão possa consultá-las, mesmo sem possuir conhecimento técnico.

Referências

- [1] JARDIM, M. E. O., E DE SOUZA, F. A. M. RioBus. 1^a Semana da Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.
- [2] COELHO, E. CPI dos Ônibus - Como funciona a CPI. <http://cpidosonibus.com.br/site/como-funciona-a-cpi.html>. Acessado em 02/11/2016.
- [3] DA SILVA JUNIOR, L. A. O movimento do código aberto. <https://www.vivaolinux.com.br/artigo/0-movimento-do-codigo-aberto>, outubro de 2009. Acessado em 26/12/2015.
- [4] HEUSSER, F. I. Understanding Open Government Data and addressing its Impact. Open Data Research Network <http://www.opendataresearch.org/reports>. Acessado em 09/11/2015.
- [5] IKEDA, A. T. E. Inovação no Serviço Público através de Dados Abertos: Um Estudo de Caso do Buus. Projeto Final, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.
- [6] OPEN KNOWLEDGE FOUNDATION. The Open Definition - Defining Open in Open Data, Open Content and Open Knowledge. <http://opendefinition.org>. Acessado em 13/10/2015.
- [7] BERNERS-LEE, T. Open, Linked Data for a Global Community. Gov 2.0 Expo, Washington, 2010.
- [8] OPEN GOVERNMENT WORKING GROUP. The Annotated 8 Principles of Open Government Data. <http://opengovdata.org>. Acessado em 13/10/2015.

- [9] RIO BUS. Rio Bus. <http://riob.us>. Acessado em 02/11/2016.
- [10] TORMENTA LABS. Rio Bus - Android Apps on Google Play. <https://play.google.com/store/apps/details?id=com.tormentaLabs.riobus>. Acessado em 02/11/2016.
- [11] RADUAN, M. A. C. Rio Bus - Acompanhe os ônibus do Rio em tempo real. <https://itunes.apple.com/br/app/rio-bus-acompanhe-os-onibus/id884914334?mt=8>. Acessado em 02/11/2016.
- [12] PREFEITURA DA CIDADE DO RIO DE JANEIRO. Dados Rio. <http://data.rio>. Acessado em 06/07/2016.
- [13] MONGODB, INC. MongoDB for GIANT Ideas | MongoDB. <https://www.mongodb.com>. Acessado em 02/11/2016.
- [14] GOOGLE INC. BigQuery - Analytics Data Warehouse | Google Cloud Platform. <https://cloud.google.com/bigquery/>. Acessado em 02/11/2016.
- [15] MOZILLA DEVELOPER NETWORK. JavaScript | MDN. <https://developer.mozilla.org/en-US/docs/Web/JavaScript>. Acessado em 02/11/2016.
- [16] NODE.JS FOUNDATION. Node.js. <https://nodejs.org/>. Acessado em 02/11/2016.
- [17] WEBB, C. Express.js and Node.js as a prototyping medium. <http://blog.mediumqualsmessage.com/understanding-expressjs-and-nodejs-as-a-medium-for-prototyping>. Acessado em 02/11/2016.
- [18] WORLD WIDE WEB CONSORTIUM - W3C. W3C HTML. <https://www.w3.org/html/>. Acessado em 02/11/2016.
- [19] VIAÇÃO IDEAL. 324 - Ribeira - Candelária. <http://www.viacaoideal.com.br/?p=502>. Acessado em 03/11/2016.

- [20] GANDHI, U. Criando Mapas de Calor - QGIS Tutorials and Tips. http://www.qgistutorials.com/pt_BR/docs/creating_heatmaps.html. Acessado em 03/11/2016.
- [21] G1. Começam em julho mudanças nas linhas de ônibus da Zona Sul do Rio. <http://g1.globo.com/rio-de-janeiro/noticia/2015/03/comecam-em-julho-mudancas-nas-linhas-de-onibus-da-zona-sul-do-rio.html>, março de 2015. Acessado em 02/11/2016.
- [22] G1. Reabertura da Av. Rio Branco altera itinerário de 180 linhas de ônibus. <http://g1.globo.com/rio-de-janeiro/noticia/2016/05/reabertura-da-av-rio-branco-altera-itinerario-de-180-linhas-de-onibus.html>, maio de 2016. Acessado em 02/11/2016.