

BRUNO DE PINHO SCHETTINO

BUS DATA ANALYSIS IN RIO DE JANEIRO CITY

Thesis presented to the Computing Graduate program of the Universidade Federal Fluminense in partial fulfillment of the requirements for the degree of Master of Science. Topic area: Software Engineering.

Advisors:

Prof. D.Sc. LEONARDO GRESTA PAULINO MURTA

Prof. D.Sc. VANESSA BRAGANHOLO MURTA

Niterói

2016

BRUNO DE PINHO SCHETTINO

BUS DATA ANALYSIS IN RIO DE JANEIRO CITY

Thesis presented to the Computing Graduate program of the Universidade Federal Fluminense in partial fulfillment of the requirements for the degree of Master of Science. Topic area: Software Engineering.

Approved on December 2016.

APPROVED BY

Prof. D.Sc. LEONARDO GRESTA PAULINO MURTA – Advisor
UFF

Prof. D.Sc. VANESSA BRAGANHOLO MURTA – Advisor
UFF

Prof. D.Sc. Marcos Lage
UFF

Prof. D.Sc. Eduardo Ogasawara
CEFET/RJ

Niterói
2016

For all the people who waste precious hours of
their lives everyday stuck in traffic jams in Rio de
Janeiro.

ACKNOWLEDGMENTS

To my parents: Sergio and Beatriz.

To my sisters: Adrianne and Ana Clara.

To my grandmothers: Ilda and Sueli.

To my grandfathers: Albertino and Zé Pinheiro.

To my uncle Roberto.

To Manolos de Nova Iguaçu: Álvaro, Bernardo, Dudu, Gabriel, Hugo, João Filipe e Neves.

To the friends I made at UFF: Abraão, Bárbara, Camila, Clarissa, Igor, Juliana, Lorena, Rodrigo, Rennan and Xuxu.

To my advisors, Leonardo and Vanessa, thank you for your support, patience and perseverance.

To the members of the committee, Marcos and Eduardo, thank you for giving your time evaluating this work.

RESUMO

Mobilidade urbana é um desafio para grandes cidades ao redor do mundo. O congestionamento é um dos problemas mais comuns em grandes cidades e tem sérias consequências para a qualidade de vida de todos os moradores urbanos. Transporte é um dos grandes desafios de qualquer cidade que tem interesse em se tornar uma Cidade Inteligente e é um componente crítico de projetos urbanos. Este trabalho tem foco na análise de dados de ônibus da cidade do Rio de Janeiro como estudo de caso. O maior objetivo é apresentar abordagens e ferramentas para facilitar o entendimento dos dados coletados, comparar diferentes momentos do tráfego e entender como o trânsito flui.

Keywords: dados de ônibus; análise; Rio de Janeiro; engarrafamento; trânsito; dados urbanos;

ABSTRACT

Urban mobility is a challenge for major cities around the world. The traffic jam is one of the most common problems in large cities and has serious consequences for the quality of life for all urban residents. Transport is one of the great challenges of any city that has an interest in becoming a smart city and is a critical component of urban design. This work focuses on analyzing bus data from the city of Rio de Janeiro as a case study. The main objective is to present the approaches and tools to make easier to understand the data collected, compare different moments of traffic and understand how the transit flows.

Keywords: bus data; analysis; Rio de Janeiro; traffic jam; transit; urban data;

LIST OF FIGURES

Figure 1. Quartiles distribution.....	25
Figure 2. Bus average speed disposal false negative example	27
Figure 3. Line bounding box example.....	28
Figure 4. Garage identification heat map	29
Figure 5. Bus statistics request in JSON format.....	30
Figure 6. Line history request in JSON format	31
Figure 7. Bus history request in JSON format.....	31
Figure 8. Buses on radius request in JSON format.....	32
Figure 9. Line bounding box request in JSON format	32
Figure 10. Line positions request in JSON format	33
Figure 11. Line stops request in JSON format	33
Figure 12. Configuration interface example.....	34
Figure 13. Line routes interface example	35
Figure 14. Positions heat map query interface.....	37
Figure 15. Positions heat map diff interface.....	38
Figure 16. Speed heat map query interface	40
Figure 17. Speed heat map diff interface.....	41
Figure 18. Loading statistics page	41
Figure 19. Valid data and disposals proportion.....	43
Figure 20. Valid data and disposals across the day	44
Figure 21. Valid data and disposals correlation	45
Figure 22. Comparison of the speed throughout the days of the week	46
Figure 23. Position diff of line 266 between average Sunday and average Friday	47
Figure 24. Speed diff of line 266 between average Sunday and average Friday	47
Figure 25. Comparison of the speed throughout the hours of the day.....	48
Figure 26. Position diff of line 266 between the average of period 1 and the average of period 4	48
Figure 27. Speed diff of line 266 between the average of period 1 and the average of period 4	49

Figure 28. (a) Tiradentes eve 2014 speed comparison with next week Sunday and average Sunday. (b) Tiradentes 2014 speed comparison with next week Monday and average Monday. (c) Tiradentes eve and Tiradentes day speed comparison on 2014.	50
Figure 29. Position diff of line 266 between the average Monday and Tiradentes 2014 holiday	51
Figure 30. Speed diff of line 266 between the average Monday and Tiradentes 2014 holiday	51
Figure 31. (a) Tiradentes eve 2015 comparison with previous week, next week and average Monday. (b) Tiradentes 2015 comparison with previous week, next week and average Tuesday. (c) Tiradentes eve and Tiradentes day speed comparison on 2015.....	52
Figure 32. (a) Labor Day eve 2014 comparison with previous week, next week and average Wednesday. (b) Labor Day 2014 comparison with previous week, next week and average Thursday. (c) Labor Day eve and Labor Day speed comparison on 2014.	53
Figure 33. (a) Labor Day eve 2015 comparison with previous week, next week and average Thursday. (b) Labor Day 2015 comparison with previous week, next week and average Friday. (c) Labor Day eve and Labor Day speed comparison on 2015.....	54
Figure 34. (a) Spain x Chile match day comparison with the average Wednesday. (b) Germany x France match day comparison with the average Friday. (c) Ecuador x France match day comparison with the average Wednesday. (d) Colombia x Uruguay match day comparison with the average Saturday. (e) Belgium x Russia match day comparison with the average Sunday. (f) World Cup Final match day comparison with the average Sunday.	56
Figure 35. Speed comparison between the months of the year	56
Figure 36. Speed comparison between each of the collected months	57

LIST OF TABLES

Table 1. API's codes and messages.....	30
Table 2. Positions heat map diff colors	37
Table 3. Speed heat map query colors.....	39
Table 4. Speed heat map diff colors	39

LIST OF ACRONYMS AND ABBREVIATIONS

GPS	– Global Positioning System
CNN	– Convolutional Neural Networks
API	– Application Programming Interface
HTTP	– Hyper Text Transfer Protocol
GUI	– Graphical User Interface
UI	– User Interface
JSON	– JavaScript Object Notation
CSV	– Comma Separated Values
XML	– Extensible Markup Language

TABLE OF CONTENTS

Chapter 1 – Introduction.....	13
1.1 Motivation	13
1.2 Goals.....	14
1.3 Organization	15
Chapter 2 – Related work	16
2.1 Methodology.....	16
2.2 Works using bus data as source.....	17
2.3 Works using other data sources	18
2.4 Conclusion.....	22
Chapter 3 – Materials and methods	23
3.1 Data collection.....	23
3.2 Data loading and cleaning strategies	23
3.2.1 Repeated records.....	24
3.2.2 Records with abnormal instantaneous speed	25
3.2.3 Records with abnormal average speed	26
3.2.4 Buses out of service	27
3.3 API.....	29
3.4 Data analysis tool.....	34
3.4.1 Configuration.....	34
3.4.2 Line routes	34
3.4.3 Positions heatmap	35
3.4.4 Speed heatmap.....	38
3.4.5 Loading statistics	40
3.5 Final remarks	41

Chapter 4 – Results.....	43
4.1 RQ1 - How reliable is the data collected from the buses without any filtering?.....	43
4.2 RQ2 - How does the traffic conditions vary in different days of the week?	45
4.3 RQ3 - How does the traffic conditions vary in different hours of the day?	47
4.4 RQ4 - Which is the behavior of the traffic on holidays compared to the average days of the week?	49
4.5 RQ5 - Which is the behavior of the traffic during a large event in the city?	54
4.6 RQ6 - How does the traffic conditions vary in different months?	56
4.7 Final remarks	57
Chapter 5 – Conclusion	58
5.1 Contributions	58
5.2 Threats to validity	59
5.3 Future work.....	59

CHAPTER 1 – INTRODUCTION

1.1 MOTIVATION

For the first time in history, more than half of the world's population lives in urban areas (FERREIRA et al., 2013). While in the recent past decision makers and social scientists faced significant constraints in obtaining the data needed to understand city dynamics and evaluate policies and practices, data are now abundant (FERREIRA et al., 2013).

According to PU et al. (2013), movement patterns are important for traffic analysts to understand the behaviors of moving objects especially in transportation management. Transport is one of the great challenges of any city that has an interest in becoming a smart city and is a critical component of urban design. Monitoring and analyzing transport data can be used to support experts in traffic on their analysis. For example, experts in transportation department can figure out why congestions are happening more frequently and find effective ways to ease the traffic load in the modern cities (PU et al., 2013). Traffic jams are one of the most common problems in large cities and have serious consequences for the quality of life for all urban residents. One effective way to understand the traffic situation and vehicle status on road networks is monitoring and analyzing the trajectory data generated from vehicles equipped with a GPS device (PU et al., 2013).

There are many kinds of vehicles that can be used as data source for a study of this kind. There are several studies using data collected from taxis as the source for studies of the traffic (FERREIRA et al., 2013; JULIANA FREIRE et al., 2014; LIU et al., 2013; PU et al., 2013; YAN et al., 2010; ZICHENG LIAO; YIZHOU YU; BAOQUAN CHEN, 2010). Some other work use private cars as their data source (ANDRIENKO; ANDRIENKO; WROBEL, 2007; GENNADY ANDRIENKO; NATALIA ANDRIENKO, 2008; YAN et al., 2010). ANDRIENKO; ANDRIENKO; WROBEL (2007) and ALBUQUERQUE et al. (2013) also use truck fleets as sources. Finally, buses are the last type of vehicle used in studies, found in the works of LÉ CUÉ et al. (2014), BARBOSA et al. (2014), ANDRADE; CRUZ (2015), MARCOS R. VIEIRA et al. (2015) and BESSA et al. (2016).

Real-time bus tracking, where available, has been well received by transit riders (THIAGARAJAN et al., 2010). Knowing where a bus is at the moment and when it will arrive in a certain position reduces the waiting time, increasing the efficiency and improving the safety and comfort of the users (THIAGARAJAN et al., 2010).

In the last years, the Rio de Janeiro city hall released the real-time GPS coordinates of all of the buses working in the city. It is possible to develop many kinds of applications with these data. For using these data with an acceptable precision, it is necessary to understand the traffic as a whole and also understand the patterns and abnormalities of the data provided. In two studies about these data release by the Rio de Janeiro city, BESSA et al. (2016) and BARBOSA et al. (2014) observed that many buses trajectories do not follow their planned route. Besides this problem, there are several issues not addressed in these works like buses out of work that keep sending their data and any failures in the GPS devices attached to the buses that can threaten the quality of the data, sending unrealistic values of position and speed of the buses.

1.2 GOALS

The goal of this work is to understand the traffic behavior patterns, using the data of the buses working in Rio de Janeiro as a proxy. This allows us to contrast the traffic in everyday situations with traffic in specific situations. Thus, we introduce the following research questions:

RQ1. How reliable is the data collected from the buses without any filtering?

This question aims to identify the kinds of abnormalities present on the collected data and the portion of these data they represent. The data remaining after filtering the abnormalities can be considered reliable to use on analysis and applications.

RQ2. How does the traffic conditions vary in different days of the week?

This question aims to study the behavior of the traffic over the week, identifying the days with more and less traffic.

RQ3. How does the traffic conditions vary in different hours of the day?

This question aims to study the behavior of the traffic over the day, identifying the hours of the day with more and less traffic.

RQ4. Which is the behavior of the traffic on a holiday compared to the average days of week?

This question aims to study the behavior of the traffic on holidays, comparing the traffic on these days with the traffic of non-holidays.

RQ5. Which is the behavior of the traffic during a large event in the city?

This question aims to study the behavior of the traffic during the 2014 Soccer World Cup, comparing the traffic on game days with the traffic of regular days.

RQ6. How does the traffic conditions vary in different months?

This question aims to study the behavior of the traffic over the months, identifying the months of the year with more and less traffic.

Aiming at answering these research questions, we conceived and implemented an approach for analyzing bus traffic data. As shown in the previous section, the data provided by the Rio de Janeiro city have some abnormalities that need to be addressed before using the data in order to make consistent analysis. For implementing this approach we developed two tools: the first one, for cleaning and loading the data to a database and the second one for visualizing the bus traffic data. Moreover, we loaded about one year of bus traffic data from the Rio de Janeiro city using these tools to allow us answering the research questions presented.

1.3 ORGANIZATION

This work is organized in five chapters including this introduction. Chapter 2 presents related work, the process and the methodology used to find them. Chapter 3 discusses the methods used in our approach, describing the tools created to collect, clean, load and analyze data. Chapter 4 exposes the results of our study, answering the research questions presented above and presents the insights of the study. Chapter 5 presents the conclusions and contributions of the study and discusses future work.

CHAPTER 2 – RELATED WORK

This chapter presents the related work and the process we followed to find them. This chapter is organized as follows. Section 2.1 describes the methodology. Section 2.2 presents the works using bus data as source. Section 2.3 shows the works using other data sources. Section 2.4 presents the conclusion of the chapter.

2.1 METHODOLOGY

The methodology we followed to find related work was divided in four steps. The first step (S1) consisted on several searches on Google Scholar. In the second step (S2) we filtered the results found on S1. In the third step (S3) we applied snowballing techniques (JALALI; WOHLIN, 2012) on the work remaining after S2. The last step (S4) consisted on a final filtering on the work remaining after S3.

The first step consisted on four searches on Google Scholar. The first search was made with the terms “urban traffic data visualization”, finding 85 papers. The second search was made with the terms “urban traffic abnormality detection”, finding 52 papers. The third search was made with the terms “urban traffic analysis”, resulting in 79 papers. The last search was made with the terms “Rio de Janeiro urban traffic”, finding 27 papers. In the end of this first step, 243 papers were found.

The second step consisted on filtering the results of S1 in two sub-steps. In the first sub-step, we filtered by title, which resulted in 76 papers. The second sub-step was filtering the remaining papers by their abstracts, resulting in 6 papers.

Snowballing is a technique for finding relevant work based on the bibliographical references of an initial set of relevant papers (backward snowballing) and on papers that cited those relevant papers (forward snowballing). This process was initiated with the 6 papers found on S2 and repeated with the new papers found on snowballing until no new relevant work is found. The papers found by the snowballing were filtered by title and abstract. In the end of this step, we had 16 selected papers.

The final filtering consists on analyzing the entire papers. After reading all the 16 studies remaining on S3, we have ended the related work research with 13 papers that were considered relevant to this work.

2.2 WORKS USING BUS DATA AS SOURCE

Using the methodology explained above, there were found 5 works using bus data as source. USapiens (MARCOS R. VIEIRA et al., 2015), Semantic Traffic Diagnosis with STAR-CITY (LÉ CUÉ et al., 2014), BusInRio (ANDRADE; CRUZ, [s.d.]), Vistradas (BARBOSA et al., 2014) and RioBusData (BESSA et al., 2016).

USapiens is a system that aims to analyze very large urban trajectory data. Its inputs are trajectory data as bus GPS data and any other datasets that can be associated with them for the analytics as rain gauge data, road incident reports and social media.

The system has two main components: Data Pre-Processing and Urban Data Analytics Modules. The first module is in charge of processing on-line GPS data and storing it in a normalized form. From the normalized data, the second module has implemented the urban data analytics functionalities. This latter module has implemented several case studies that focused mainly on quality of bus service in a city.

The pipeline of USapiens first cleans, normalizes and integrates the data, and then implements the case studies. For this work, the case studies are: bus uniformity analysis; bus route verification; traffic flux analysis; bus travel time analysis of variance; and bus arrival time prediction.

Semantic Traffic Diagnosis with STAR-CITY describes a study of the traffic in the cities of Dublin, Bologna, Miami and Rio de Janeiro using the STAR-CITY system. The study focus is different in each city.

In Dublin the study is focused on diagnosing traffic congestion using traffic accidents, road works and social events data as sources of explanation. In Bologna the study is directed to diagnosing bus congestion using road works data as sources of explanation. In Miami the study focuses on diagnosing bus bunching which refers to a group of two or more buses that were scheduled to be evenly spaced running along the same route, instead running in the same location at the same time. In Rio the study is focused on diagnosing low on-time performance, such as buses that are heavily delayed.

BusInRio presents a mobile app developed to explore the bus data provided by the Rio de Janeiro City. The mobile app has two visualizations. In the first one it is possible to choose a bus line and see all the buses working on that line in real-time. In the second one it is possible to see a heat map of the current traffic situation in the whole city.

The architecture is divided in four parts: the web crawler that is responsible to clean the data and populate the databases; the API which is responsible to query the database and

respond to external HTTP requests; the mobile app that makes requests to the API and shows the data in a map; and the databases, which are responsible to store the data loaded by the web crawler.

Vistradas is a Web-based visualization system that allows users to visualize use cases related to trajectories of public vehicles. Being aware of the possible existence of abnormal records, they implemented a pre-processor that handles problems such as: no information about the direction in which the bus is traveling; and, in some cases, wrong latitude/longitude positions and poor time resolution. After the cleaning, the pre-processor also does the data fragmentation and normalization to build cumulative space-time bus movements.

The use cases were: analysis of bus uniformity, which aims to detect buses of the same line running very close to each other; verification of bus route, which aims to verify if the buses are traveling in their expected routes; and impact of events in bus traffic, which aims to identify if events like road constructions, natural phenomena, sport or music events, among others, can cause a negative impact on the city's traffic. In the analysis of bus uniformity use case they have detected buses of the same line very close to each other on the rush hours. In the verification of bus route they could detect a few GPS entries far away from the expected route, indicating a possible detour. In the impact of events in bus traffic use case they detected a decrease of 8 km/h on the bus average speed after the Perimetral overpass fall.

RioBusData is a visual analytics system that aims to automatically identify anomalous trajectories in bus routes using Convolutional Neural Networks (CNN) (LECUN et al., 1998). In this system, any trajectory that contains at least one GPS entry that does not follow the expected behavior in space (e.g., buses outside their planned routes) and time (e.g., delayed buses) is considered as an outlier.

In their experiment, they used the data collected from the buses running in Rio de Janeiro city from September 26, 2013 to January 9, 2014, totaling over 151 million records. The applications of the system were analyzing the outliers in this dataset as a whole and separated by spatial outliers and temporal outliers. Using RioBusData is also possible to detect which lines and hours of the day have more and less outliers.

2.3 WORKS USING OTHER DATA SOURCES

Using the methodology explained above, there were found 8 relevant works using other data sources. CubeView (SHEKHAR et al., 2002), AITVS (LU; BOEDIHARDJO; ZHENG, 2006), Visual Analytics Tools for Analysis of Movement Data (ANDRIENKO et al., 2007), A proactive application to monitor truck fleets (ALBUQUERQUE et al., 2013), T-

Watcher (PU et al., 2013), Automatic construction and multi-level visualization of semantic trajectories (YAN et al., 2010), A Visual Analytics System for Metropolitan Transportation (LIU et al., 2011) and Visual exploration of big spatio-temporal urban data (FERREIRA et al., 2013).

CubeView is a web-based visualization package for observing rapid summarization of major traffic trends. The data flow is as follows. The basic map and raw data are cleaned, transformed, and loaded into the data warehouse module, which provides the multidimensional views and the Online Analytical Processing operations for data visualization as well as a variety of data mining analysis tools such as classification, clustering and outlier detection. The discovered patterns or rules are then visually displayed as maps or charts for further interpretation.

The system has a Graphic User Interface (GUI) that draws a highway map using the geographic coordination information of each station. It accepts queries from users and sends queries to a middle tier, which further requests the traffic data from the database tier. The GUI can display traffic video, detect outlier stations, and show highway volume maps for a user-specified time, date, and highway stations.

AITVS is a web based traffic visualization system that provides visualization components to analyze and monitor traffic conditions. The system presents information in various formats to observe and analyze traffic trends.

AITVS architecture is divided in three components: the visualization engine; the web server; and the database server. The system updates its traffic data once every minute when it receives real-time data from the loop detectors, but can be adjusted to aggregates of any specified time value.

The system provides six visualization components that allow the users to see several metrics of a roadway system. All the visualizations show three kinds of data: speed, volume, and occupancy. The visualizations are as follows: Time Plot that shows the three kinds of data of a particular station for a specified period of time of day; Date Plot that shows the three kinds of data of a particular station for a specified date range; Highway Station Plot that shows the three kinds of data of a particular station for a specified set of highway station nodes; Highway Stations vs. Time Plot that shows the three kinds of data of a set of station nodes for a specified time of day; Highway Stations vs. Day of the Week Plot that shows the three kinds of data of a set of station nodes for a specified time of day; and Time vs. Day of Week Plot that shows the three kinds of data of all days in the week for each time of day.

Visual Analytics Tools for Analysis of Movement Data describes a flexible movements analysis framework that can be used to analyze several kinds of moving data such as private cars and trucks. The framework is divided into four modules: data preprocessing module; extraction of significant places module; extraction of trips module; and the examination of trips module.

The data preprocessing module adds extra fields to the received movement data such as the time interval and the distance in space to the next position. The second task of the data preprocessing module is to filter the data to remove sequences of records corresponding to absence of movement. The extraction of significant places module runs a clustering algorithm that detects places where the records are stopped as a person work place or depots where the truck's loads are taken. The extraction of trips module identifies movement of an object from one significant place to another. The examination of trips module is a set of visualization tools that helps the users to view individual trips, clustering of trips, and summarization of trips and examination of visited places.

A proactive application to monitor truck fleets monitors truck trips and traffic-related facts about road conditions that may affect the traffic on the routes of these trips. They use workflows to model truck trips and traffic-related tweets to collect real-time facts about road conditions.

The solution that they implemented consists on 4 modules. The first one stores data about the truck fleet and geographic data as street/road maps and truck routes. The second one obtains traffic-related messages from Twitter, transforms these messages into structured data and geo-references these data. The third module monitors truck trajectories. The last module analyzes the geo-referenced structured dynamic data to detect traffic-related facts of interest, determines which facts may affect planned truck trips and alerts the user about such facts.

T-Watcher is an interactive visual analytics system for monitoring and analyzing complex traffic situations in big cities via taxi trajectory data for regions, roads, and vehicles. The system architecture consists of three primary components: a data preprocessing module; a visualization rendering module; and a user interaction module.

T-Watcher uses preprocessed GPS data as input and has three main components: the region fingerprint, which displays the whole spatial temporal distribution and uses a ring-map-based radial layout design to show historical data; the road fingerprint, which can further analyze locations selected by the user from the region fingerprint; and the vehicle fingerprint, which helps the user to explore historical statistical information while monitoring the real-time situation.

Automatic construction and multi-level visualization of semantic trajectories demonstrates SeMiTri, which is a system for automatic construction and multi-level visualization of semantic trajectories from raw mobility traces. The architecture of the system is divided into four layers: Trajectory Computation Layer that performs preliminary works such as trajectory data cleaning, dividing a trajectory into several episodes, and computing structural-level trajectories; Semantic Annotation Layer which annotates trajectories with data of semantic places available from 3rd party geographic sources; Semantic Trajectory Analytics Layer that computes additional statistical information (e.g. distribution values such as mean, variance, max, min) for trajectories; and the Web Interface which presents users with a visual and integrative way to query and retrieve the enriched semantic trajectories at different abstracted levels.

The system has several visualization capabilities such as: Spatio-Semantic Trajectories that demonstrates multiple levels of trajectory data abstraction, showing raw GPS tracks, raw trajectories, structured trajectories (e.g. stops/moves), and semantic trajectories (e.g. home-office-supermarket-home); User Interactions that provides a Web interface to query and visualize trajectories at different abstraction levels; and Analytics Results that highlights statistical analytics results of semantic trajectories like the average speed when the user is moving.

A Visual Analytics System for Metropolitan Transportation describes VAST, a system developed to study large scale transportation data integrating visualization and data analytics methods using taxi trips of China as the data source. VAST provides a history scheme for users to track their operations so they can always go back to any step in their operations.

On VAST's GUI the query result area shows results in text whereas the statistics area in bar charts, pie graphs, and other visual analytical representations are shown when it is applicable. There are several predefined queries as finding the location or trajectory of a given taxi for a given time or a given time period, retrieving the taxi distribution in the city for a given time period, given the start and destination locations, finding the fastest and the shortest routes, among others.

Visual exploration of big spatio-temporal urban data, uses data of taxi trips in the city of New York to understand the city dynamics. They build a system called TaxiVis, which provides a visual query model that supports spatio-temporal queries over origin-destination data. This visual model allows users to manipulate the data using graphical widgets and visualize their results. The visual model implemented support several classes of queries, as the three classes proposed by Peuquet (1994): identify a set of objects at a given location and

time; given a time and a set of objects, describe the locations occupied by them; and describe the times a set of objects occupied a given set of locations.

The authors investigate three case studies: investigating taxi activity in different regions; exploring movement; and studying behavior over time. In the use case directed to analyze the different regions, the analyses showed that Midtown and the Upper East Side are the most active areas of New York. They also noticed that over the weekend the number of drop-offs in Downtown is increased. In the use case aiming to explore the movement, the authors compared trips starting at the airports with trips starting at the train stations at New York. They could detect a number of pick-ups much higher around the train stations than on the airports. In the last use case, with the purpose of studying behavior over time, results indicate that the number of taxis on the streets could potentially be reduced on the Memorial Day holiday. They were also able to detect a large decrease of the number of taxis during the week of Hurricanes Sandy and Irene.

2.4 CONCLUSION

The related works cover a wide range of features and provide many kinds of analysis. However, just a few of them discuss their outlier detection and data cleaning methods. The ones that describe their outlier detection and data cleaning methods do not cover all the needs for cleaning the data provided by the Rio de Janeiro city hall. Moreover, the visualization systems studied presents complex UIs and do not have enough options for filtering and comparing the desired data.

CHAPTER 3 – MATERIALS AND METHODS

To answer the research questions presented in the introduction of this work and to help us to understand the traffic behavior in Rio de Janeiro, we have created tools to collect, clean, load and analyze data. This chapter presents the materials and methods used in this work. It is organized as follows. Section 3.1 introduces the data collected from the buses working on Rio de Janeiro city. Section 3.2 shows our strategies for data loading and cleaning the collected data. Section 3.3 presents the API we created to ease the access of the data stored in the database. Section 3.4 shows the tools created for visual analysis. Lastly, section 3.5 contains the final remarks of this chapter.

3.1 DATA COLLECTION

For this work, we used the data provided by the Rio de Janeiro City Hall through their data.rio Web portal (PRJ, 2016). This Web Portal provides a JSON file that is generated every minute with the most updated information about the position, speed and line of every bus working in Rio de Janeiro city. Since a new file is available every minute in the same URL, if we just read the information and do not store it somehow, we would not be able to access that information later, since it would have been replaced by new data. For this reason, we have created a script that requests and stores the JSON file with the most updated information every minute. In the end of the day, all the JSON files are compressed in a ZIP file named with the date timestamp. These ZIP files are available in our server (GEMS, 2014). Besides the buses positions, the data.rio server also provides one CSV file for each line containing the GPS coordinates of the line stops and another CSV file for each line containing the line itinerary. Since this information does not change frequently, we only requested and stored them once.

3.2 DATA LOADING AND CLEANING STRATEGIES

After collecting the data, we have chosen to load it into a database, so we can easily perform queries on the data. Since the ZIP files mentioned in the previous section are available in our server, we can download them and read the JSON files to load their data to an object-relational database with support to spatial data so we can search the desired data using SQL.

Each record inside the JSON file contains six fields: *timestamp*, indicating the date and time the data was collected; *bus number*, determining from which bus the data came; *line*

number, establishing the line route that the bus is working; *latitude* and *longitude*, which together designate the bus position; and *speed*, informing the instant speed of the bus at the moment of collection.

It was developed a tool that analyses each record of these JSON files and loads them to the database. This process of data loading does not have any automation yet. It is necessary to download the ZIP files and run the loading tool every time you want to load new data.

After some analysis, we noticed that some data appeared to be abnormal. To handle situations like this, we created a new table called Disposals. This table keeps all the records considered abnormal, while the Bus Positions table keeps all the records considered normal. The following subsections presents the abnormality detection algorithms used in the cleaning process.

3.2.1 REPEATED RECORDS

At first, we noticed that some records were duplicated. The duplication of the records can cause problems, for instance, when we want to calculate the speed average of the buses. If a record is duplicated, the average would consider the same information twice and the result would be wrong. The most common situation found in the data collected is that a new record comes with exactly the same timestamp of the last stored record of that same bus. We also noticed cases in which the new record comes with a timestamp previous than the timestamp of the last record for that same bus. We noticed that in this last situation, 99% of the records timestamp are between the last ten records of the same bus.

To improve performance in this cleaning strategy, we had to fetch the last ten records of each bus from the database and store them in the RAM memory before beginning to load new data. In the other 1% of the records coming with a timestamp older than the last record's, we would have to query the database to be certain that this new record is not duplicated. Because of performance reasons we chose to not perform this query. If there is no record of the same bus with the same timestamp in the database, the new record may be inserted in the Bus Positions table depending on the other cleaning strategies results. Otherwise, it is inserted in the Disposals table with the disposal reason "Repeated record".

In this case, we do not have false positives. All the records of this kind inserted in the Disposals table are duplicated for sure. We can have false negatives if the position came with a timestamp prior to the last 10 records.

3.2.2 RECORDS WITH ABNORMAL INSTANTANEOUS SPEED

Besides the problem of the duplication of the records, we also noticed that some records had above normal speeds. For instance, we found speeds up to 1256 kilometers per hour. A faulty GPS device is probably the cause of values that high. As in the case of the record duplication, records with abnormal instantaneous speed would also cause errors when we want to calculate a speed average, for instance.

To discover which is the maximum value of speed we can consider as normal, we populated the database with 5 weeks of data, having almost 293 million records and used an interquartile range method to detect the outliers. The lower quartile $Q1$ is the lower value that is higher than 25% of the sorted data set. The higher quartile $Q3$ is lower value that is higher than 75% of the sorted dataset. Figure 1 shows a chart representing the quartile distribution. The outlier was calculated using the formula $Q3 + 1.5*(Q3-Q1)$, and the outlier threshold found was 85.57 km/h.

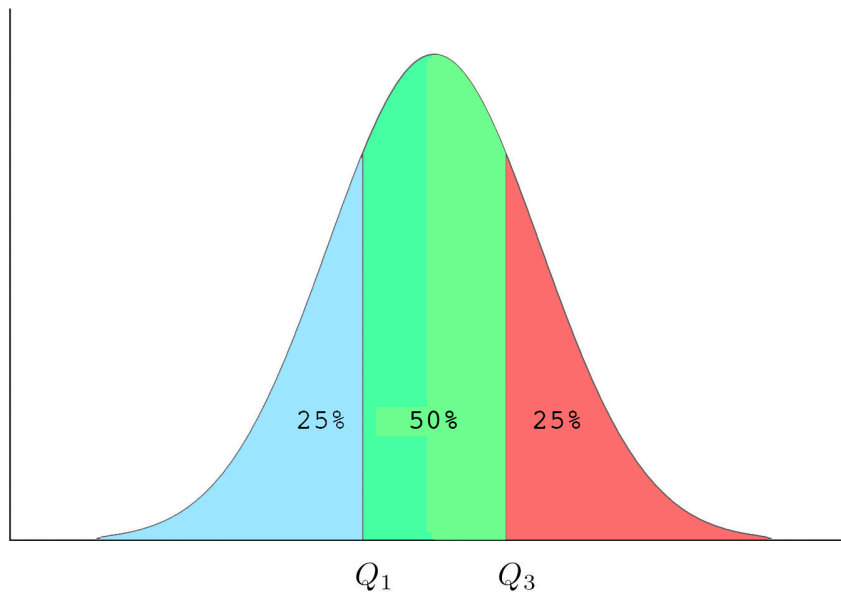


Figure 1. Quartiles distribution

This way, when we are inserting a new record, we take a look at the provided instantaneous speed and if it is above 85.57 km/h, the record is inserted in the Disposals table with the disposal reason “Instantaneous speed higher than 85,57 km/h”. Otherwise, the record may be inserted in the Bus Position table depending on the results of the other filters.

In this case, we can have false positives and false negatives. The false positives are those in which the GPS device is working fine and the bus is really traveling in a speed higher than 85.57 kilometers per hour. On the other hand, this filter can have false negatives when it

misses speed errors when the GPS is not working properly, but it is informing lower speeds than it should.

3.2.3 RECORDS WITH ABNORMAL AVERAGE SPEED

Another problem was detected when we noticed that some consecutive records of the same bus have abnormal distance. For instance, we found two consecutive records in which the distance was over 600 kilometers and the timestamps difference was only a few minutes. This error is probably also caused by a defective GPS device.

To avoid this kind of error, when we are inserting a new bus position, we compare the position of the new record with the position of the last inserted record of the same bus. If the distance in a straight line between the position of the new record and the position of the last one indicates that the bus is traveling with average speed higher than 85.57 km/h (the outlier presented on section 3.2.2), this new record is inserted in the Disposals table with the disposal reason “Average speed higher than 85,57 km/h”. Otherwise, it may be inserted in the Bus Positions table depending on the other cleaning strategies result.

In this case we can also have false positives and false negatives. The false positives are those in which the bus is really traveling in a high speed and the average speed is in fact higher than 85.57 km/h. The false negatives can happen when the two positions are closer than the real path used by the bus. Figure 2 shows an example of a false negative situation. The red marker represents the first record position and the blue marker represents the second record position. The red line is the straight-line distance between the two record’s positions, which is shorter than the path that the bus actually traveled, represented by the blue line (the blue line represents the bus itinerary provided by data.rio). This difference can cause our algorithm to recognize the average speed as a value lower than 85.57 km/h, when the actual average speed could be much higher.



Figure 2. Bus average speed disposal false negative example

3.2.4 BUSES OUT OF SERVICE

We noticed that some buses do not turn off the GPS sensor when they are inactive, such as when they are standing in the garage. We could notice several records in sequence, in which a bus had zero instantaneous speed and only a few meters traveled. Even when the bus is standing in the garage, the subsequent records may have few meters distance between them due to the attached error in GPS devices (BRADFORD W. PARKINSON; JAMES J. SPILKER, 1996). This kind of problem would cause errors in some kinds of analysis, as the other problems shown above. To eliminate this problem, we perform three complementary methods.

The first one checks if the record came with the line number. We notice that several records came with no line number assigned, which leads us to believe that they are not working. This way, when a record came without a line number, it is inserted in the Disposals table with the disposal reason “Record without line”. In this case, we can have false positives when the bus is actually working and the record comes without a line number assigned.

The second method is to set a bounding box limited by the line itinerary coordinates and assume that all the records outside this bounding box are out of order. It was used a configurable tolerance in the bounds to prevent discarding records with GPS precision errors and small detours. We could use another strategy in which we only assume that a bus is active when it is passing exactly through the streets that compose the route. Using that strategy, we would discard records in which the bus is going through a small detour, maybe caused by a construction work or any event occurring on the street that it should pass. Also, knowing if the bus is traveling exactly on the streets of the route would cause performance issues. For

these reasons, we choose to use the bounding box strategy mentioned above. Figure 3 shows an example where the line path is represented in dark blue, and the bounding box is represented in light blue.

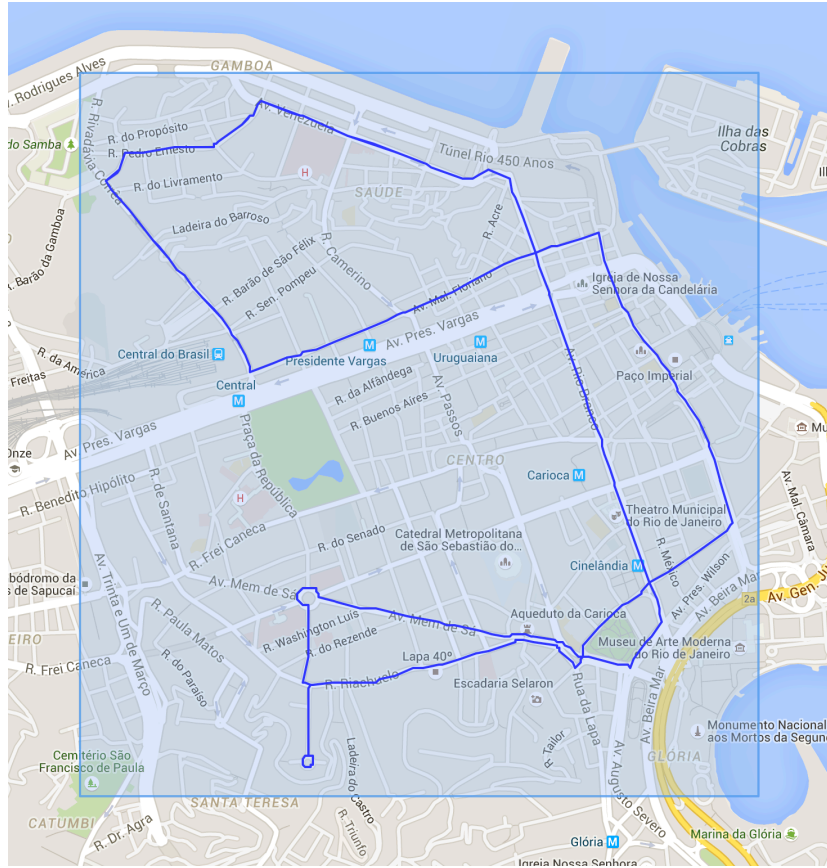


Figure 3. Line bounding box example

All the records of this line with position inside this light blue box are inserted in the Bus Positions table, and all the records with position outside of the box are inserted in the Disposals table with the disposal reason “Bus out of route”. In this case, we can have false positives and false negatives. The false positives are those in which the bus makes big detours outside the bounding box when it is actively working on the line. The false negatives occur when the garage is located inside the bounding box.

The last method is to identify the buses stopped at the garages. For doing this, we used the same database from Section 3.2.2. Figure 4 shows the heat map generated with data between 2 AM and 3 AM. This hour range was chosen because it is the time when, despite a few exceptions, the buses are not working. Then, most of the buses are likely be parked at the garages of their companies. For identifying the garages, we made a manual check of each of the clusters in the image, looking at Google Maps Street View images in the area and searching for information on the websites of the companies. It was created a bounding box around every garage identified.

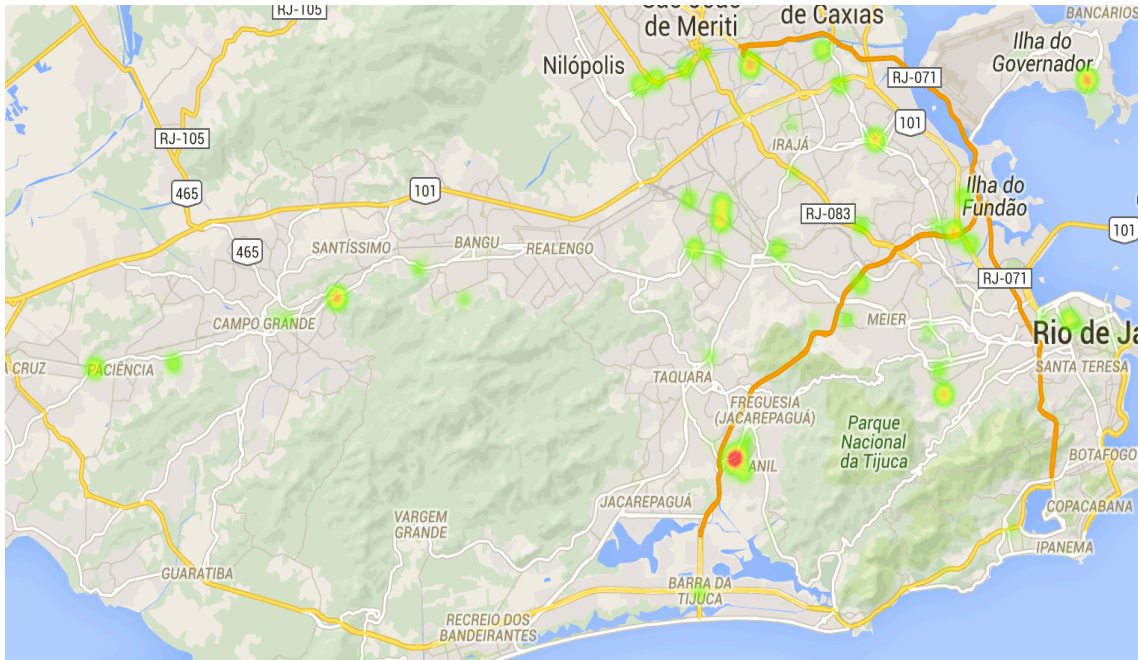


Figure 4. Garage identification heat map

When a record is inside one of these bounding boxes, it is inserted in the Disposals table with the disposal reason “Bus at the garage”. Otherwise, it may be inserted in the Bus Positions table depending on the other cleaning strategies results.

In this case we can have only false negatives. The false negatives can happen when the garage is located inside the line bounding box. In this case it is not possible to track when the bus is parked at the garage using this method.

3.3 API

The database filled with the data presented above can be useful to many kinds of applications. It is necessary to ease the data access for people unfamiliar with object-relational databases with spatial objects support. For this reason we created a HTTP API, which is a well-known and easy to use tool to retrieve data for the user. Our API comprises several request types that it can respond to in XML and JSON formats. The available requests types are: bus statistics; bus valid data; line history; bus history; buses on radius, line bounding box; line stops; and line positions. All the API’s requests return the fields: *code*; *message*; and other fields needed to represent the requested data itself. The field code is filled with a number that informs the response status and the field message is the explanation of this code. Table 1 shows the codes and it’s respective messages. The data field has different structures in each request, and is presented next.

Table 1. API's codes and messages

Code	Message
0	Success
1	Unexpected error: <Error message>
2	Missing parameters: <Missing parameters list>
3	Line not found
4	Bus not found

The first request type, bus statistics, returns the bus average speed and record average count grouped by time and day of a week and can be filtered by a single line. If the line is not provided, the request will return with code 2 and a message indicating that the line was not provided. Figure 5 shows an example of a successful request from line 371 in JSON format.

```
Request:
GET
http://localhost:3000/api/v1/bus_positions/statistics?line=371&format=json

Response:
{
  "code": 0, "message": "Success", "line": "371",
  "data": {
    "monday": {
      "speed": [ avg_speed_at_midnight, ... ,
                 avg_speed_at_11PM ],
      "count": [ avg_count_at_midnight, ... ,
                 avg_count_at_11PM ],
    },
    "tuesday": {
      "speed": [ avg_speed_at_midnight, ... ,
                 avg_speed_at_11PM ],
      "count": [ avg_count_at_midnight, ... ,
                 avg_count_at_11PM ],
    },
    "...": {
      "speed": [ avg_speed_at_midnight, ... ,
                 avg_speed_at_11PM ],
      "count": [ avg_count_at_midnight, ... ,
                 avg_count_at_11PM ],
    },
    "sunday": {
      "speed": [ avg_speed_at_midnight, ... ,
                 avg_speed_at_11PM ],
      "count": [ avg_count_at_midnight, ... ,
                 avg_count_at_11PM ],
    }
  }
}
```

Figure 5. Bus statistics request in JSON format

The next request type, line history, returns all the bus positions data of a specific line. If the line is not provided, as in the bus statistics request, this request will return with code 2 and a message indicating that the line was not provided. Figure 6 shows an example of a request from line 371 in JSON format.

```

Request:
GET
http://localhost:3000/api/v1/bus_positions/line_history?line=371&format=json

Response:
{
  "code": 0, "message": "Success", "line": "371",
  "columns": ["longitude", "latitude", "speed", "bus_number", "time"],
  "data": [
    ["-43.342476", "-22.883194", "29", "C51634",
    "2015-02-20 12:43:29"],
    ["-43.187298", "-22.906137", "14", "C51637",
    "2015-02-20 12:43:54"],
    ["-43.209293", "-22.910532", "43", "C51603",
    "2015-02-20 12:43:57"]
  ]
}

```

Figure 6. Line history request in JSON format

The bus history request type returns all the bus positions data of a specific bus. If the bus is not provided, this request will return with code 2 and a message indicating that the bus was not provided. Figure 7 shows an example of a request from bus C51637 in JSON format.

```

Request:
GET
http://localhost:3000/api/v1/bus_positions/bus_history?bus=C51637&format=json

Response:
{
  "code": 0, "message": "Success", "line": "371",
  "columns": ["longitude", "latitude", "speed", "line_number", "time"],
  "data": [
    ["-43.187298", "-22.906137", "14", "371", "2015-02-20 12:43:54"],
    ["-43.193111", "-22.905512", "0", "371", "2015-02-20 12:56:57"],
    ["-43.193859", "-22.905773", "0", "371", "2015-02-20 12:57:58"]
  ]
}

```

Figure 7. Bus history request in JSON format

The buses on radius request type returns all the bus positions data near a radius from the latitude and longitude position provided in the request. If the latitude, longitude or the radius is not provided, this request will return with code 2 and a message indicating that some of the parameters were not provided. Figure 8 shows an example of a request from latitude -22.906137, longitude -43.187298 and radius 10 meters in JSON format.

```

Request:
GET
http://localhost:3000/api/v1/bus_positions/on_radius?lat=-
22.906137&long=-43.187298&rad=10&format=json

Response:
{
  "code": 0, "message": "Success",
  "columns": ["longitude", "latitude", "speed", "bus_number", "line_number",
    "time"],
  "data": [
    ["-43.187332", "-22.906086", "0", "C51512", "371",
      "2015-02-20 12:46:12"],
    ["-43.187298", "-22.906137", "14", "C51637", "371",
      "2015-02-20 12:43:54"],
    ["-43.187223", "-22.906112", "20", "C51637", "371",
      "2015-02-20 12:47:51"]
  ]
}

```

Figure 8. Buses on radius request in JSON format

The line bounding box request type returns the max and min latitude and longitude that compose the bounding box of the line. If the line is not provided, this request will return with code 2 and a message indicating that the line was not provided. Figure 9 shows an example of a request from line 371 in JSON format.

```

Request:
GET
http://localhost:3000/api/v1/line_bounding_box/from_line?line=371&format=json

Response:
{
  "code": 0, "message": "Success",
  "columns": ["longitude", "latitude", "speed", "bus_number",
    "line_number", "time"],
  "data": {
    "max_lat": -22.875, "min_lat": -22.9117,
    "max_long": -43.1852, "min_long": -43.354299999999995
  }
}

```

Figure 9. Line bounding box request in JSON format

The line positions request type returns all the positions that compose the line route. If the line is not provided, this request will return with code 2 and a message indicating that the line was not provided. Figure 10 shows an example of a request from line 371 in JSON format.

```

Request:
GET
http://localhost:3000/api/v1/line_positions/from_line?line=371&format=json

Response:
{
  "code": 0, "message": "Success",
  "data": [
    { "line_number": "371", "latitude": -22.9052,
      "longitude": -43.1877, "sequence_number": 0,
      "description": "371-PRACA SECA X PRACA DA REPUBLICA",
      "company": "Fetranspor" },
    { "line_number": "371", "latitude": -22.9057,
      "longitude": -43.1875, "sequence_number": 1,
      "description": "371-PRACA SECA X PRACA DA REPUBLICA",
      "company": "Fetranspor" },
    ...,
    { "line_number": "371", "latitude": -22.9077,
      "longitude": -43.19, "sequence_number": 774,
      "description": "371-PRACA SECA X PRACA DA REPUBLICA",
      "company": "Fetranspor" }
  ]
}

```

Figure 10. Line positions request in JSON format

The line stops request type returns all the positions that contains stop points in the line route. If the line is not provided, this request will return with code 2 and a message indicating that the line was not provided. Figure 11 shows an example of a request from line 371 in JSON format.

```

Request:
GET
http://localhost:3000/api/v1/line_stops/from_line?line=371&format=json

Response:
{
  "code": 0, "message": "Success",
  "data": [
    { "line_number": "371", "latitude": -22.9019,
      "longitude": -43.3478, "sequence_number": 1,
      "description": "371-PRACA SECA X PRACA DA REPUBLICA",
      "company": "Fetranspor" },
    { "line_number": "371", "latitude": -22.9056,
      "longitude": -43.1875, "sequence_number": 2,
      "description": "371-PRACA SECA X PRACA DA REPUBLICA",
      "company": "Fetranspor" },
    ...,
    { "line_number": "371", "latitude": -22.9014,
      "longitude": -43.3457, "sequence_number": 77,
      "description": "371-PRACA SECA X PRACA DA REPUBLICA",
      "company": "Fetranspor" }
  ]
}

```

Figure 11. Line stops request in JSON format

3.4 DATA ANALYSIS TOOL

To help us to identify problems like those presented on Section 3.2 and understand the transit in the Rio de Janeiro city, we created an analysis tool. A good way to view information about bus positions is showing the corresponding records on a map. The visual analysis is divided in three subsections: Section 3.4.1 shows the configurations interface; Section 3.4.2 introduces the line routes interface; Section 3.4.3 shows the positions heat map interface; Section 3.4.4 presents the speed heat map interface; and Section 3.4.5 shows the loading statistics interface.

3.4.1 CONFIGURATION

To allow the users to adapt the tools for their own analysis it was created the configuration interface, which allows the user to customize all the parameters used to display the data in the analysis tool. Figure 12 shows an example of parameters configuration.

The screenshot shows a web-based configuration interface titled "Configuration". It is divided into four main sections:

- Positions heatmap:** Contains four input fields labeled "Range 1" through "Range 4" with values -15, -5, 5, and 20 respectively.
- Speed diff heatmap:** Contains four input fields labeled "Range 1" through "Range 4" with values -20, -5, 5, and 20 respectively.
- Speed query heatmap:** Contains three input fields labeled "Range 1", "Range 2", and "Range 3" with values 15, 30, and 60 respectively.
- Other:** Contains two input fields: "Bounding box gap" with value 0.001 and "Search radius" with value 15.

An "Update" button is located at the bottom right of the configuration area.

Figure 12. Configuration interface example

3.4.2 LINE ROUTES

The line routes interface allows the user to choose a bus line by it's number, and visualize the route made by the buses operating in this line. A switch can be turned on or off to show or hide the stop points of the line and the bounding box presented in the previous section. Figure 13 shows an example of the line positions interface, in which we can see the

route of line 10 marked out by the blue line and its stops indicated by the markers. The numbered blue circles are bus stops clusters that indicate several stops very next to each other in the line segment.

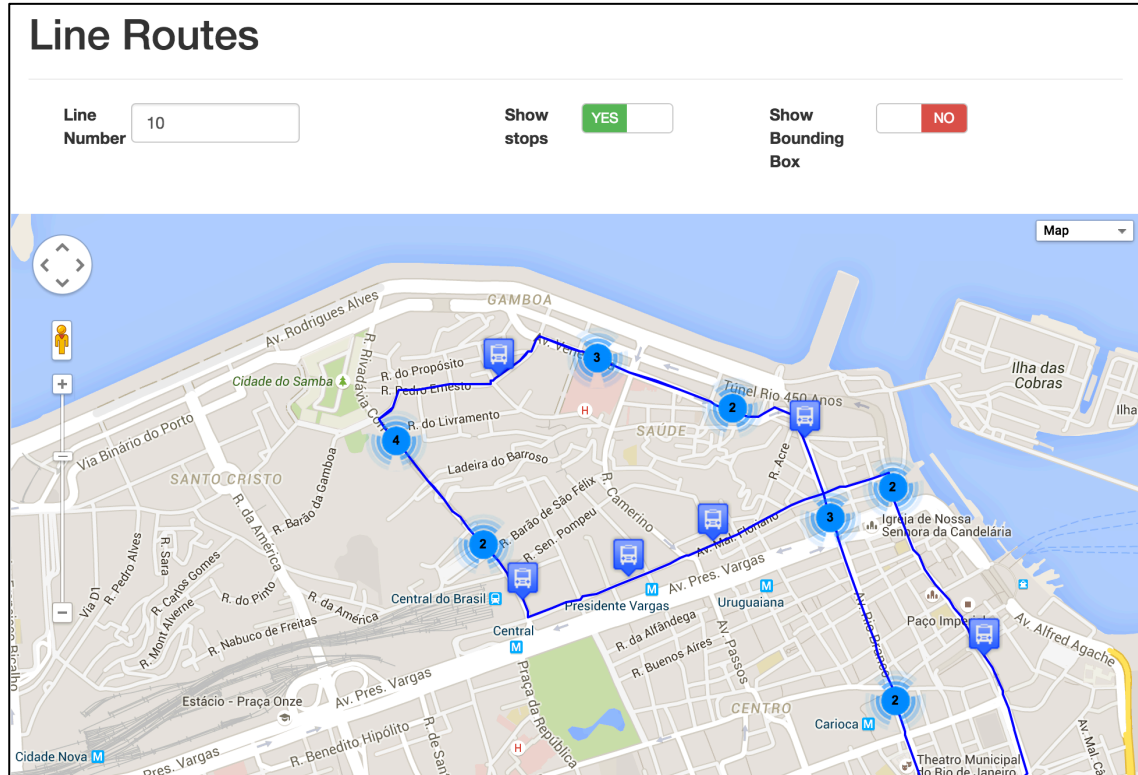


Figure 13. Line routes interface example

3.4.3 POSITIONS HEATMAP

In the positions heatmap interface it is possible to choose a bus line by its number, and visualize a heatmap composed by the positions of the buses operating in this line. As in the line positions interface, a switch can be turned on or off to show or hide the stop points of the line. This interface can show two types of heatmaps: one for a single query result and another one for comparing two different query results. The user can alternate between the query tab, which shows the result of a single query and the diff tab, which shows the result of the comparison of two different query results.

The query heatmaps have also other 3 filtering options besides the line option: by the time of the day; by the day of week; and by a specific date. These filtering options can be combined with the line number filtering option to show a different heatmap for each query result chosen by the user.

The first one is filtering by the time of the day. As the traffic can vary greatly during the day, this kind of filtering is useful for enabling the user to see the heatmap generated in

different moments of the day. For instance, it is expected to be more buses operating in the range of 6 AM to 10 PM than at other times.

The second filtering option is the day of the week. As in the time of day option, this filtering option was made because the traffic can vary greatly from one day of the week to another. For instance, it is expected that the traffic is more intense in Fridays than on Sundays. The last filtering option is the date. This filtering shows the heat map of a specific day. For instance, the user can choose this option to see the heat map of a day when a big accident has occurred. This option can be combined with the time of the day option but cannot be combined with the day of the week. This restriction was designed to avoid incompatibility errors. For instance, if this combination was allowed, the user could pass January 1, 2014 (which was a Wednesday) as the date parameter, and pass Sunday as the day of week parameter. So, when combining both filters, the result would be empty.

In the query tab, the results of filtering are shown using a Heat Map. In this visualization, the red color is used to paint the areas with the highest number of points, the green color is used to paint the areas with the lowest number of points and the intermediate colors are used to paint areas with intermediate number of points. Thus, areas with fewer points are painted with a color closest to green and the ones with more points are painted with a color closest to red. Figure 14 presents an example of the positions heat map interface, in which we can see the areas that have higher and lower number of bus positions of the line 10, in Tuesdays, between 12 and 7 PM.

In the diff tab, we can compare two query results and see a heat map with the between them. The points outside of the line route are discarded because it is not trivial to calculate the difference between distant points. So, for each point inside the line route, compare the amount of bus positions near it in both query results and calculate the between the two values in percentage. To draw the colors, we preferred to use the Maps Polylines feature rather than the Heat Map feature because it is easier to fill the line route and draw the specific color that we want.

Table 2 shows the meaning of the five colors used to represent the degree of difference between the query results.

Figure 15 shows an example of the positions heat map diff interface, in which we can see the comparison between the line 10, in Tuesdays, between 12 and 7 PM and the same line in Thursdays, at the same time range. The gray segments indicates that the numbers of positions of the buses working on the line 10, in Tuesdays, between 12 and 7 PM are close to the number of positions of the buses working on the line 10, in Thursdays, between 12 and 7

PM. The light red and dark red segments indicate that the second query result values are lower than the first query results in that segment. The light green and dark green segments indicate that the second query result values are higher than the first query results in that segment. The ranges that control the colors in this heat map can be adjusted in the configuration interface.

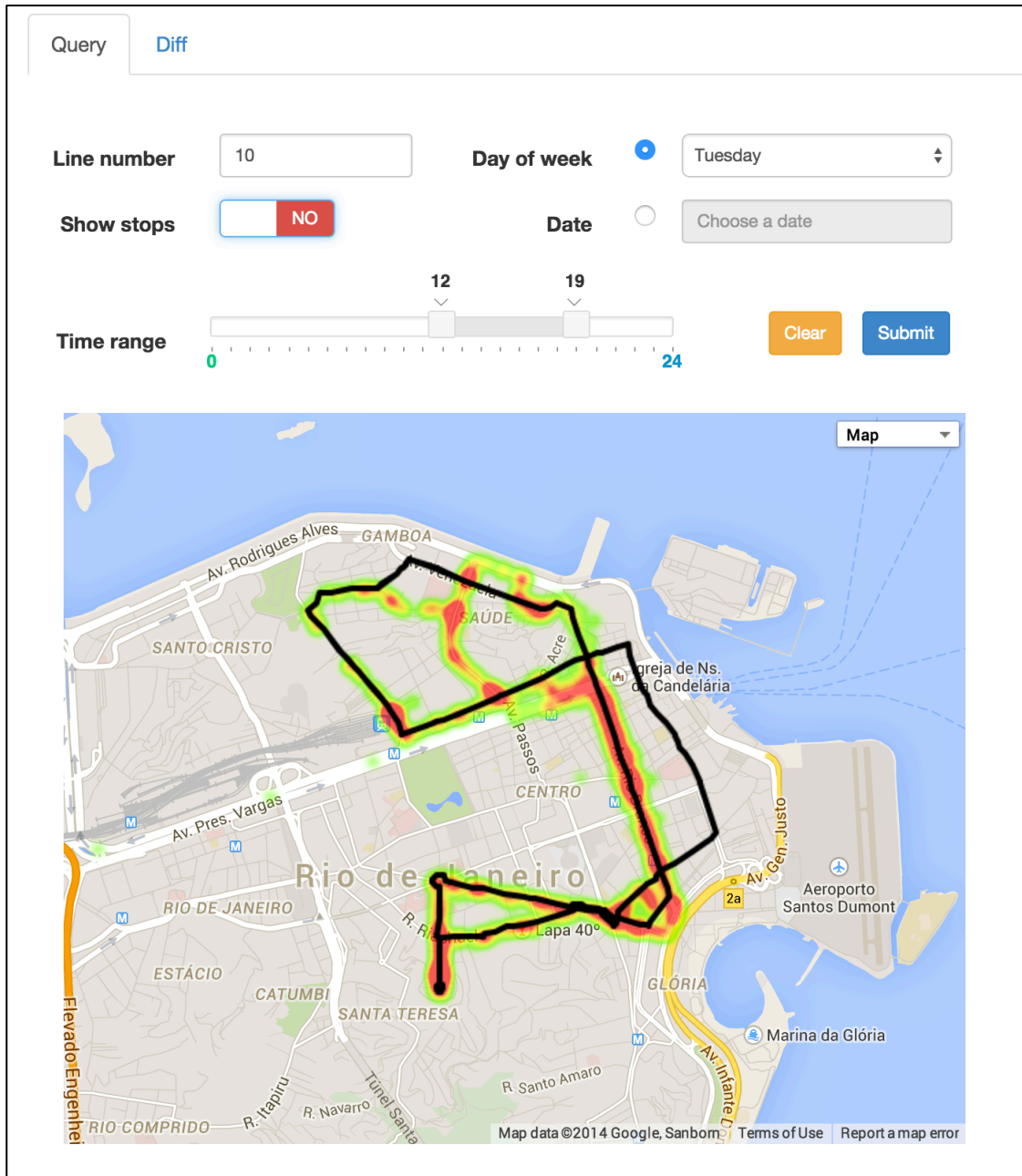


Figure 14. Positions heat map query interface example

Table 2. Positions heat map diff colors

Color	Message
Light red	The difference between the base query result and the second query result is higher than positions heat map range 4
Dark red	The difference between the base query result and the second query result is between

	positions heat map range 4 and positions heat map range 3
Grey	The difference between the base query result and the second query result is between positions heat map range 3 and positions heat map range 2
Dark Green	The difference between the base query result and the second query result is between positions heat map range 2 and positions heat map range 1
Light Green	The difference between the base query result and the second query result is lower than positions heat map range 1

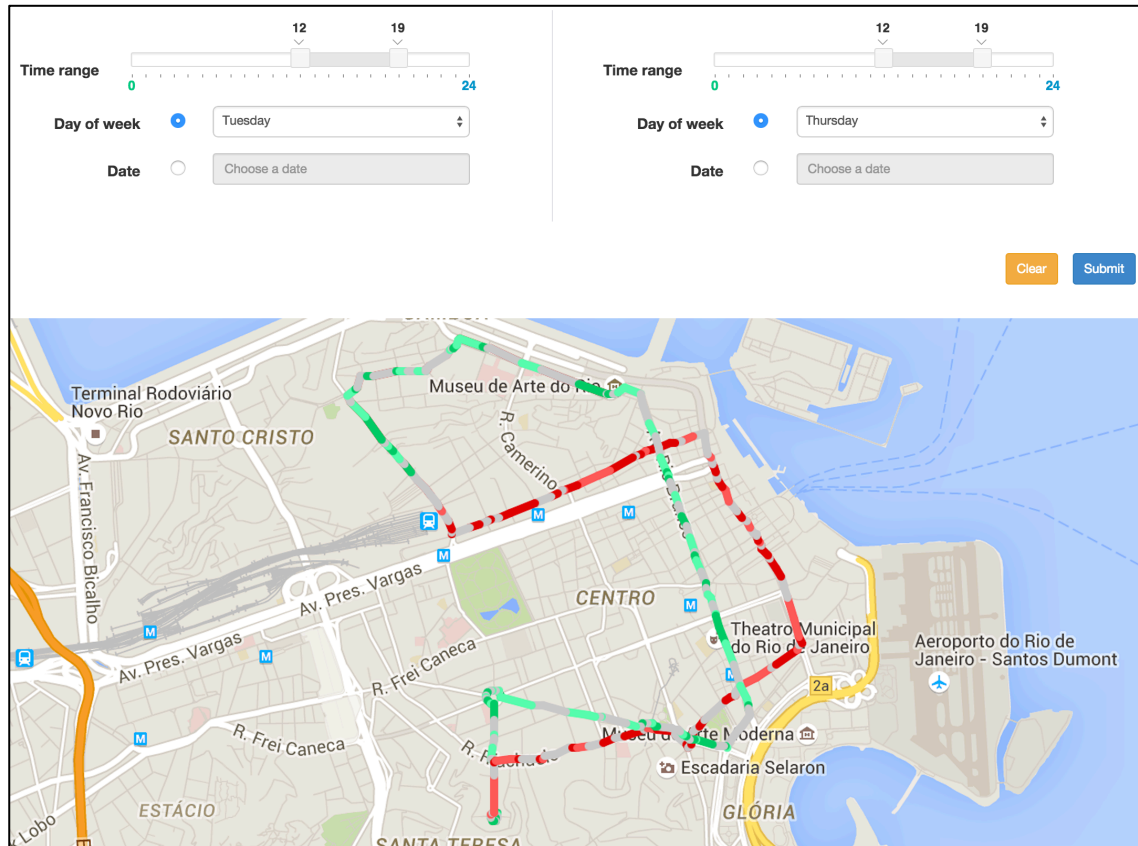


Figure 15. Positions heat map diff interface

3.4.4 SPEED HEATMAP

In the speed heat map interface is possible to choose a bus line by its number, and visualize a heat map composed by the speed of the buses operating in this line. This interface has the same filtering options presented in the positions heat map and it also has the query and diff tabs.

In both, query and diff tabs, the results of filtering are shown using Google Maps Polyline for the same reason we used it in the positions heat map. In the query visualization, we pick each point inside the line route and calculate the speed average in a configurable radius near it.

We have four colors to represent the speed intervals. Figure 16 shows an example of the speed heat map query interface, in which we can see the speed average through the line 10, in Tuesdays, between 12 and 7 PM.

Table 3. Speed heat map query colors

Color	Message
Light red	The speed from the query result is higher than speed query heat map range 3
Dark red	The difference between the base query result and the second query result is between speed query heat map range 3 and speed query heat map range 2
Dark Green	The difference between the base query result and the second query result is between speed query heat map range 2 and speed query heat map range 1
Light Green	The difference between the base query result and the second query result is lower than speed query heat map range 1

In the diff tab, we can compare two query results and see a heat map with the distinction between them. For each point inside the line route, we compare the speed average near it in both query results and calculate the difference between the two values, subtracting the second value from the base value.

Table 4 shows the meaning of the five colors used to represent the degree of difference between the query results.

Table 4. Speed heat map diff colors

Color	Message
Light red	The difference between the base query result and the second query result is higher than speed diff heat map range 4
Dark red	The difference between the base query result and the second query result is between speed diff heat map range 4 and speed diff heat map range 3
Grey	The difference between the base query result and the second query result is between speed diff heat map range 3 and speed diff heat map range 2
Dark Green	The difference between the base query result and the second query result is between speed diff heat map range 2 and speed diff heat map range 1
Light Green	The difference between the base query result and the second query result is lower than speed diff heat map range 1

Figure 17 presents an example of the speed heat map diff interface, in which we can see the comparison between the line 10, in Tuesdays, between 12 and 7 PM and the same line in Thursdays, at the same time range.

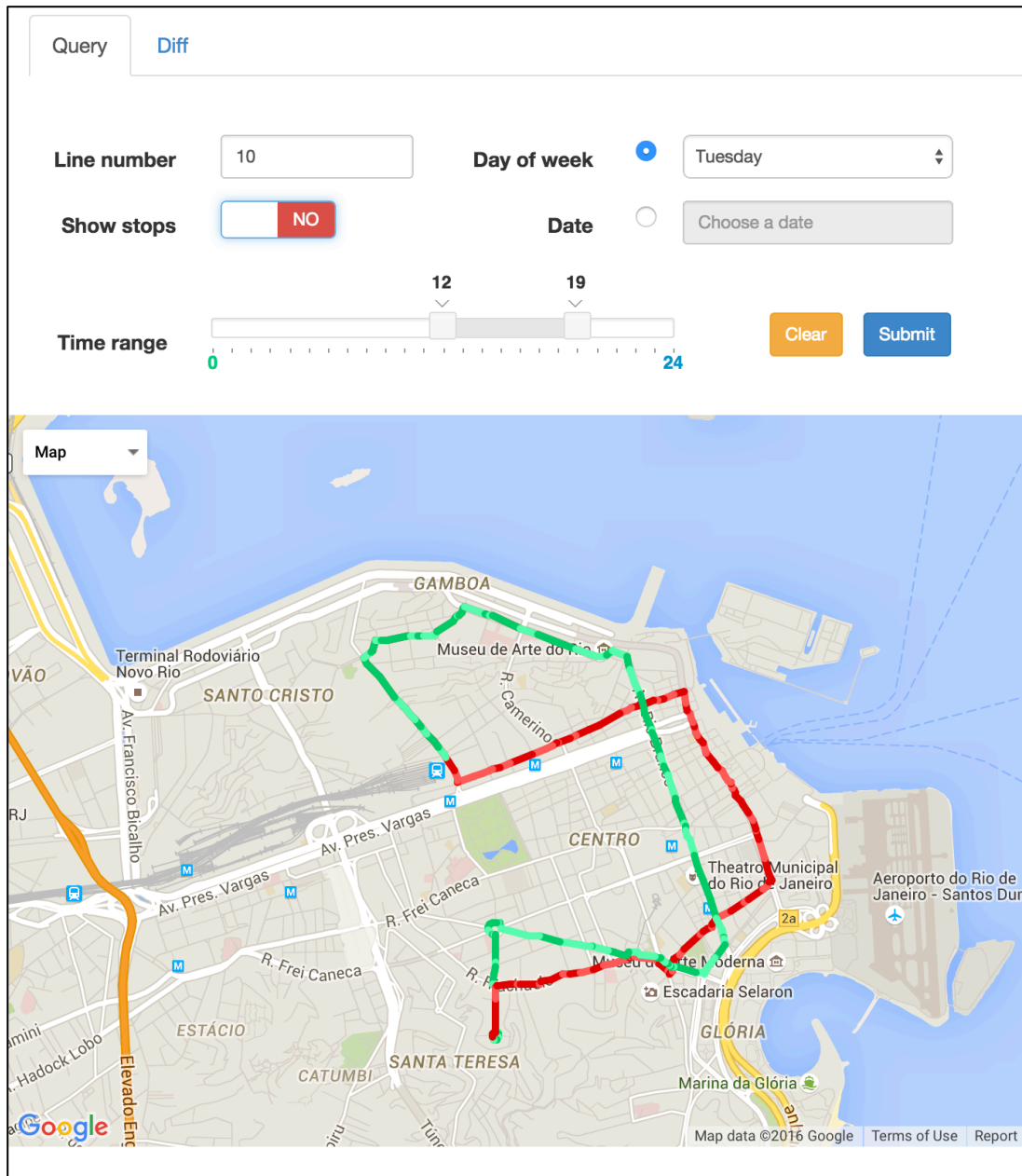


Figure 16. Speed heat map query interface

3.4.5 LOADING STATISTICS

In the loading statistics page we can see how many valid records were loaded to the database as valid data and how many were loaded as disposal by each reason. Figure 18 shows the chart that can be seen in the loading statistics page.

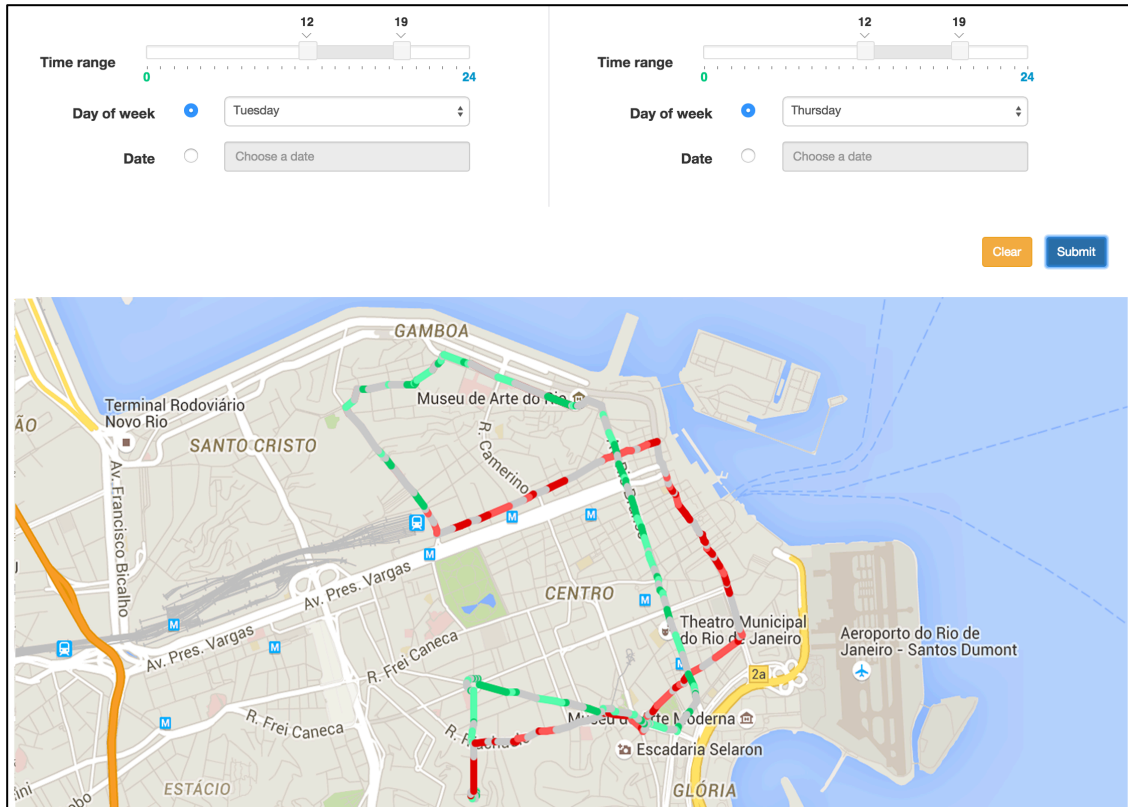


Figure 17. Speed heat map diff interface

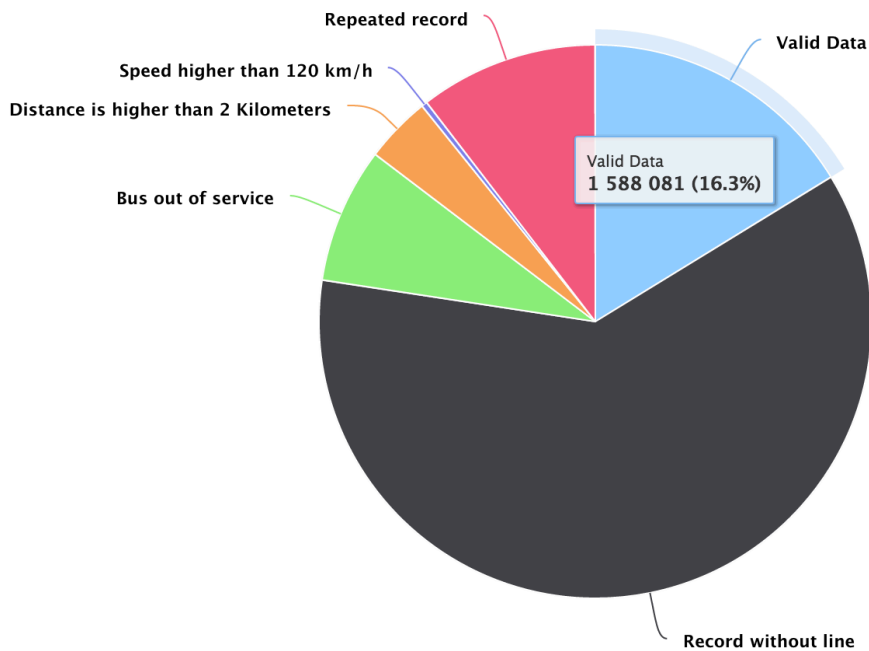


Figure 18. Loading statistics page

3.5 FINAL REMARKS

In this chapter we have presented the materials and methods used in this work, like the approaches and tools for loading, cleaning and analyzing the data. In the next chapter we

present the results of the study made using these approaches and tools shown in this chapter to answer the research questions presented in the introduction of this work.

CHAPTER 4 – RESULTS

This chapter aims to answer the research questions presented in the introduction of this work and it is organized as follows. Section 4.1 answers RQ1: “How Reliable is the data collected from the buses without any filtering?”, Section 3.1 answers RQ2: “How does the traffic conditions vary in different days of the week?”, Section 4.3 answers RQ3: “How does the traffic conditions vary in different hours of the day?”, Section 4.4 answers RQ4: “Which is the behavior of the traffic on a holiday compared to the average days of week”, Section 4.5 answers RQ5: “Which is the behavior of the traffic during a large event in the city?” and Section 2.1 answers RQ6: “How does the traffic conditions vary in different months?”.

4.1 RQ1 - HOW RELIABLE IS THE DATA COLLECTED FROM THE BUSES WITHOUT ANY FILTERING?

To answer this question we loaded all the data collected from the city hall website from February 20th 2015 to March 26th 2015. This resulted in a database with 302,498,658 position records, including valid data and disposals. As shown on Figure 19 only 32% of the loaded data is considered valid. The figure shows 29% of disposals by records without line, 13% of repeated records, 11% of records out of the line itinerary, 10% of records stopped at the garage, 5% of records with average speed higher than 85,57 km/h and less than 1% of records with instantaneous speed higher than 85,57 km/h.

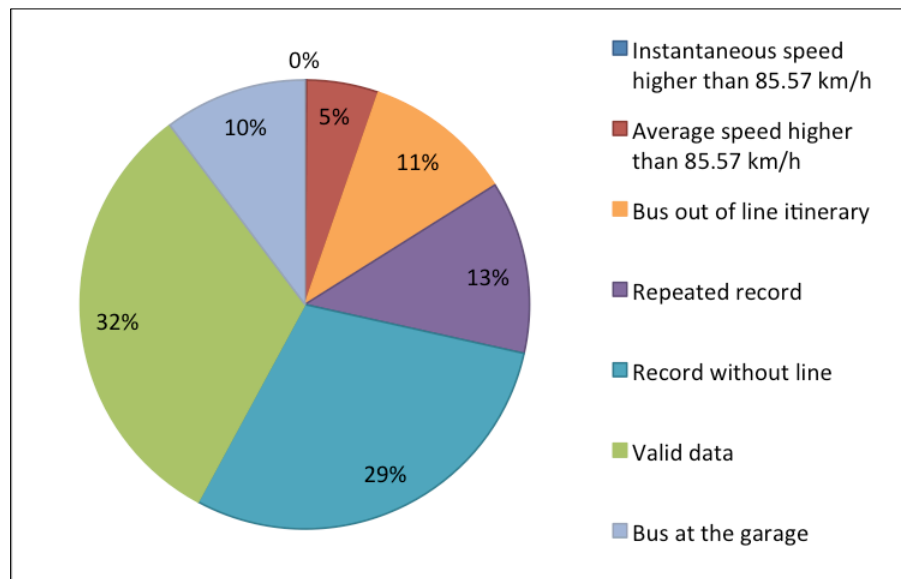


Figure 19. Valid data and disposals proportion

It is visible on Figure 20 that the distribution of valid data and disposals throughout the day. Except the disposals by instantaneous speed higher than 85.57 that have a very small

number of records across all the day, we can split the day in 3 time ranges to understand better our analysis: from 5 AM to 10 AM, when people start going to work and the buses start to move; from 10 AM to 8 PM, when the life in the city is more active; and from 8 PM to 5 AM, when most people already got back home from work.

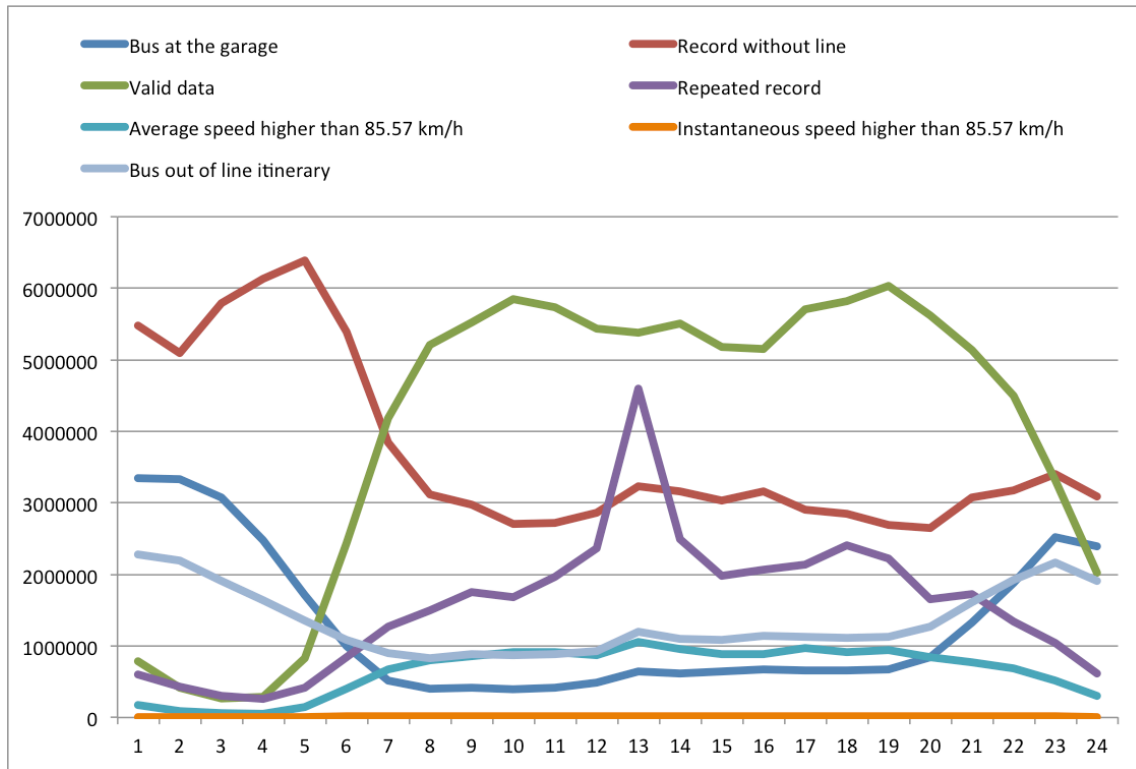


Figure 20. Valid data and disposals across the day

In the first time range, between 5 AM and 10 AM, the numbers of valid data; disposals by average speed higher than 85.57 km/h; and repeated records are increasing. On the other hand, the disposals by record without line; bus at the garage; and bus out of line itinerary present the opposite behavior in this same time range.

In the second time range, between 10 AM and 8 PM, the valid data and all disposals do not vary much, except for the disposals by repeated record that have a peak around 13 PM.

In the last time range, between 8 PM and 5 AM, the numbers of valid data; disposals by average speed higher than 85.57 km/h; and repeated record are decreasing. Moreover, the disposals by record without line; bus at the garage; and bus out of itinerary are increasing.

Figure 21 shows a correlation table where the blue circles indicate that the correlation is positive and the red circles point that the correlation is negative. The darker the shades of red and blue, the stronger the correlations are.

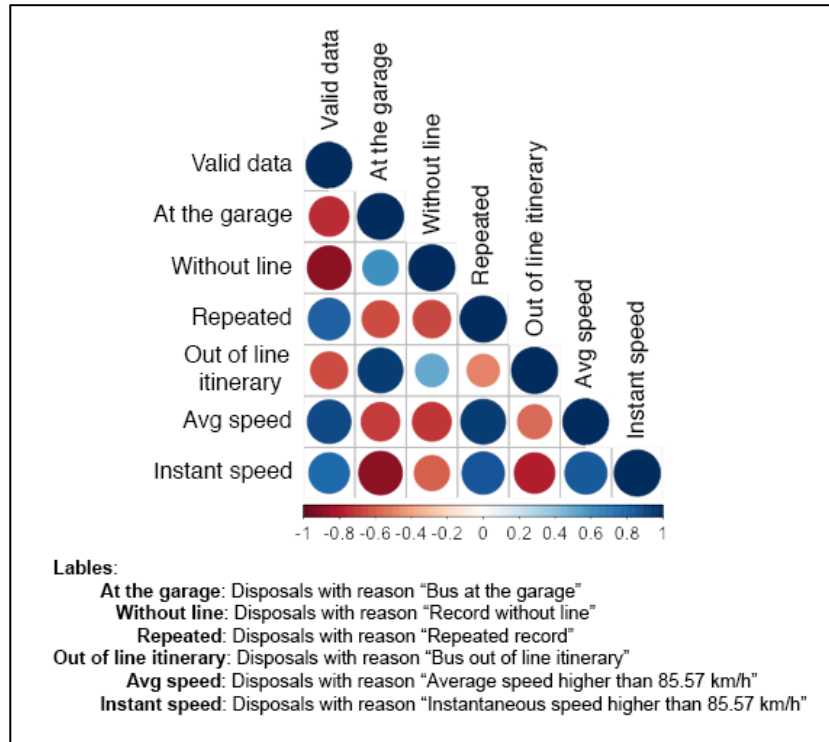


Figure 21. Valid data and disposals correlation

The numbers of valid data; disposals by average speed higher than 85.57 km/h; disposals by instantaneous speed higher than 85.57; and disposals by repeated record have a positive correlation. This occurs because these four kinds of data seem to point to the same thing: the buses are working.

The disposals by record without line; bus at the garage; and bus out of line itinerary also have a positive correlation. The same way these three disposals seem to point to one thing: the buses are not working.

The correlations between the data in the first and second group are negative values. This fact confirms that these two groups have opposite meanings.

Only 32% of the data would be reliable without any of the presented filters.

4.2 RQ2 - HOW DOES THE TRAFFIC CONDITIONS VARY IN DIFFERENT DAYS OF THE WEEK?

To answer this question and the following questions we loaded all the data collected from the city hall website from April 2014 to October 2015. We only saved the valid data on the database in order to reduce the occupied disk space. In the end of the load we had 1,253,724,203 records taking over 340 GB of disk space.

Figure 22 shows the instantaneous speed of the buses over the days of the week. Sunday is the day with higher average speed, followed by Saturday. Since most people do not work on weekends, the traffic is better in these days as expected. All the weekdays have very similar behaviors, having a barely visible decrease of speed over the week. Although the traffic is expected to be worse on Fridays since many people travel on Fridays to enjoy the weekend, the speeds on Fridays are very similar to the other weekdays in our study.

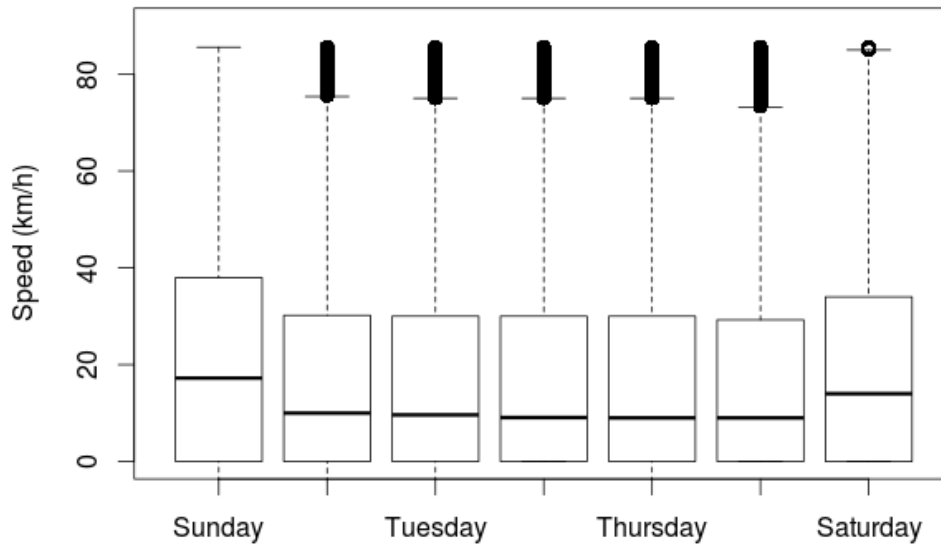


Figure 22. Comparison of the speed throughout the days of the week

Figure 23 shows the position diff of the line 266 between average Sunday and average Friday, generated by our position heat map diff tool. Line 266 was chosen because it's itinerary passes both through the city commercial center and living neighborhoods. It is visible that the majority of the heat map is filled with green. This indicates that most of the line itinerary has more buses working on this line on Fridays than on Sunday.

Figure 24 shows the speed diff of line 266 between average Sunday and average Friday, generated by our speed heat map diff tool. It is visible that most of the heat map is filled with red. This indicates that the traffic is predominantly worse on this line on Friday than on Sunday.

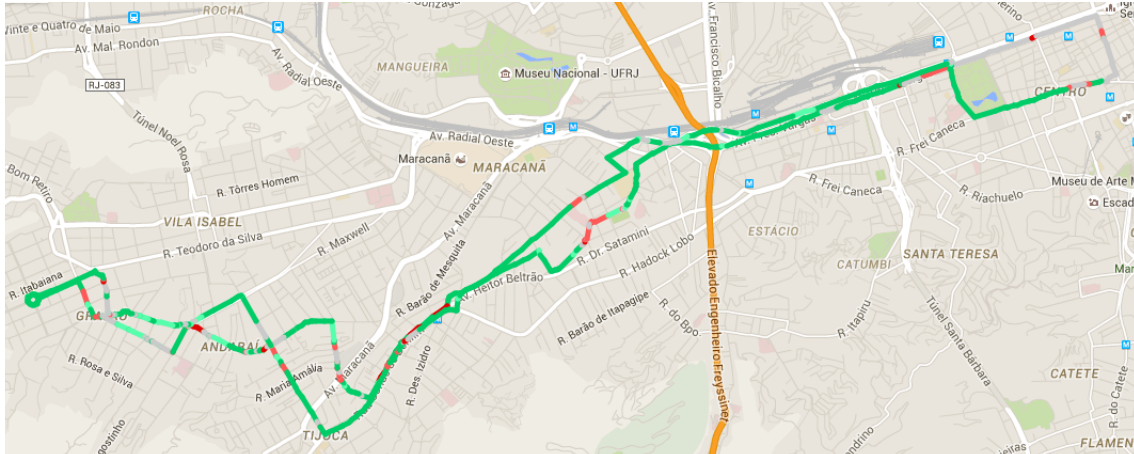


Figure 23. Position diff of line 266 between average Sunday and average Friday

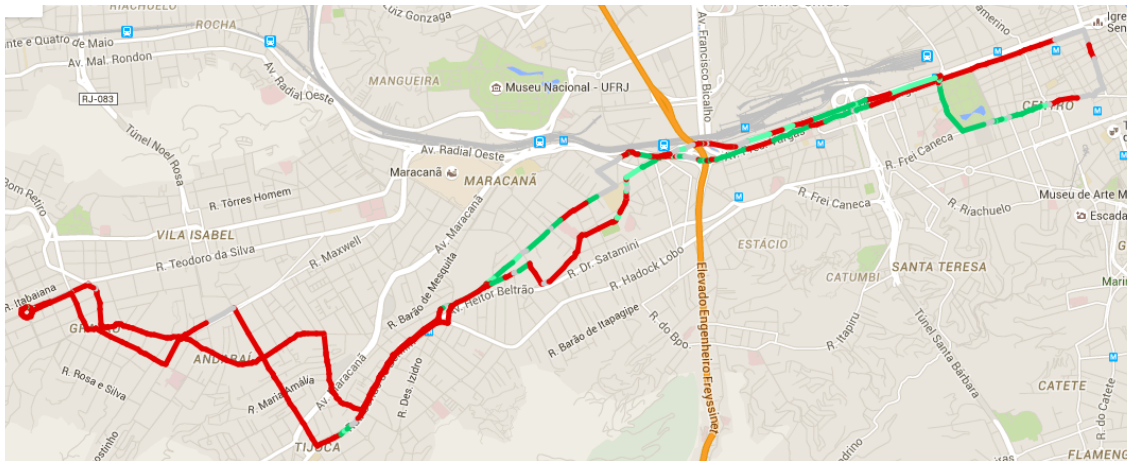


Figure 24. Speed diff of line 266 between average Sunday and average Friday

The traffic presents similar behavior over the weekdays and it is considerably better on weekends.

4.3 RQ3 - HOW DOES THE TRAFFIC CONDITIONS VARY IN DIFFERENT HOURS OF THE DAY?

To answer this question, we picked all the data in our database and divided it in hours of the day. Figure 25 shows a comparison of the speed throughout the hours of the day. Between midnight and 4AM, named *period 1*, the speed has considerable variations and remains high. The low movement of people in this time range can explain the high values. The speeds have consecutive considerable decreases between 4AM and 7AM. We name this as *period 2*. The movement of people going to work and the increase of the number of buses working in this time range can explain the decreases. From 7AM until 3PM, named *period 3*, the speeds have consecutive slight decreases. The constant and stable movement of people in the city can explain the stability of the speeds in this time range. From 3PM until 6PM, named *period 4*,

the speeds have consecutive considerable decreases. The movement of people going home from work is the probable cause of this effect. As most people are going home at the same time in this time range, traffic jams are common. The speeds have consecutive considerable increases between 6PM and midnight, named *period 5*. As most of the people go home around the same time, in this time range the speeds tend to grow as people arrive home.

Figure 26 shows the position diff of the line 266 between the average of *period 1* and the average of *period 4*, generated by our position heat map diff tool. It is visible that the majority of the heat map is filled with green. This indicates that most of the line itinerary has more buses working on this line on *period 4* than on *period 1*.

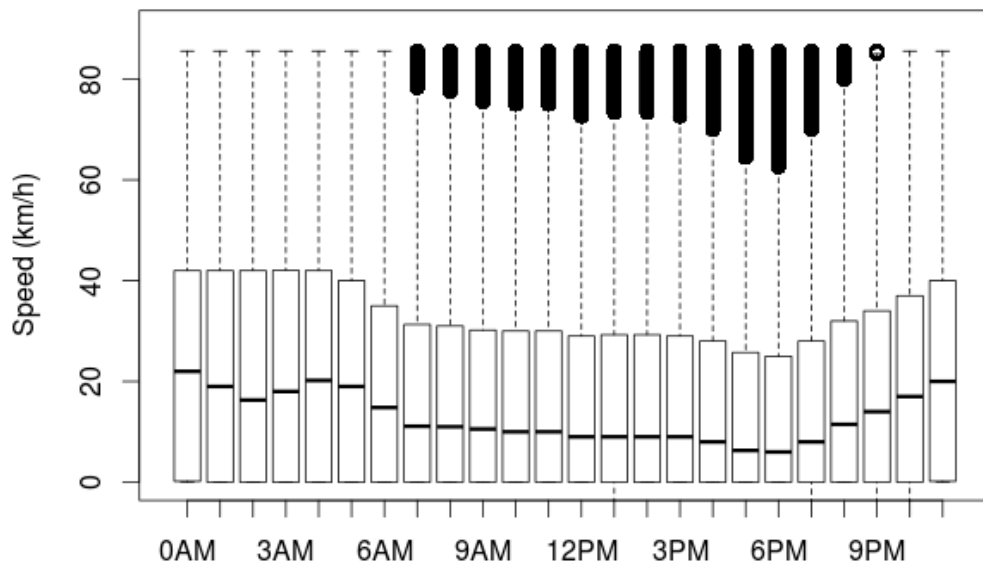


Figure 25. Comparison of the speed throughout the hours of the day

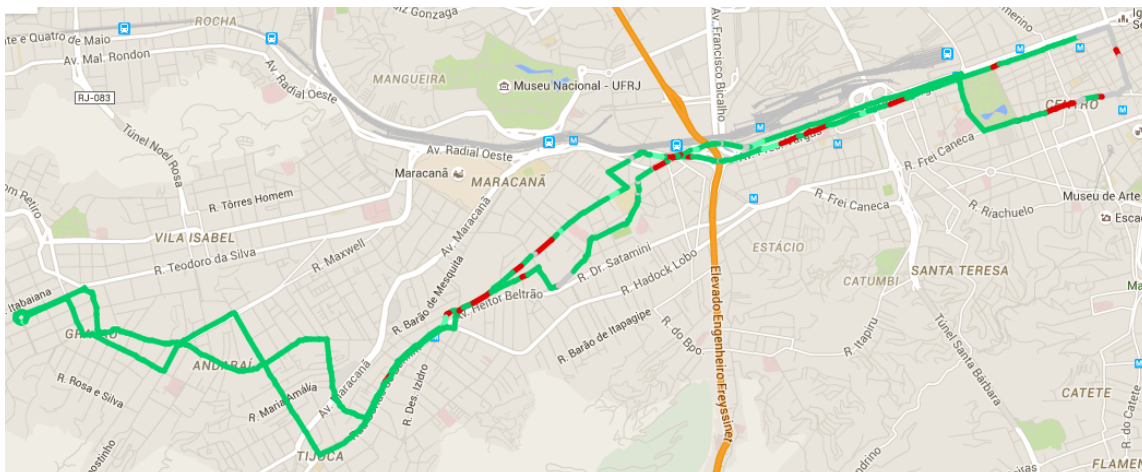


Figure 26. Position diff of line 266 between the average of period 1 and the average of period 4

Figure 27 shows the speed diff of line 266 between the average speed of *period 1* and the average speed of *period 4*, generated by our speed heat map diff tool. It is visible that most of

the heat map is filled with red. This indicates that the traffic is predominantly worse on this line on *period 1* than on *period 4*.

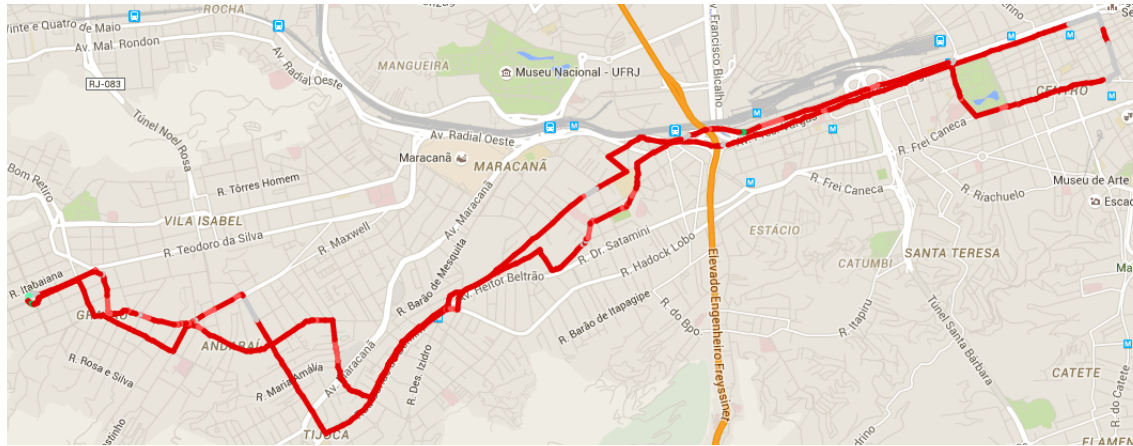


Figure 27. Speed diff of line 266 between the average of period 1 and the average of period 4

The traffic presents high variation through the day. The range with lower speed is the end of the afternoon hours and beginning of the night hours. The range with higher speed is the early morning hours.

4.4 RQ4 - WHICH IS THE BEHAVIOR OF THE TRAFFIC ON HOLIDAYS COMPARED TO THE AVERAGE DAYS OF THE WEEK?

To answer this question we chose two holidays on 2014 and 2015: Tiradentes, which is celebrated on April 21th; and Labor Day, which is celebrated on May 1st. Tiradentes day was a Monday on 2014 and a Tuesday on 2015. Labor Day was a Thursday on 2014 and a Friday on 2015.

Figure 28 (a) presents the speed comparison between Tiradentes eve (a Sunday), next week Sunday and the average Sunday. We could not compare it with the week before Tiradentes day and Tiradentes eve on 2014 because we started collecting the data on April 16th, 2014. It is visible that the speed is slightly higher on Tiradentes eve than on next week Sunday and considerably higher than on the average Sunday. As seen on Section 3.1, the traffic on weekends is better than on weekdays. For this reason, the holiday eve traffic is just a little better than the average of the equivalent day of week.

Figure 28 (b) presents the speed comparison between Tiradentes (a Monday), next week Monday and the average Monday. It is visible that the speed is much higher on Tiradentes day than on next week Monday and the average Monday. In this case, the traffic was better on a holiday than on the average of the equivalent day of week.

Figure 28 (c) shows the speed comparison between Tiradentes eve and Tiradentes day on 2014. It is visible that the speed is slightly higher on Tiradentes eve than on Tiradentes day. In this case, both holiday eve and holiday day have similar behaviors and the traffic is slightly better than on weekends.

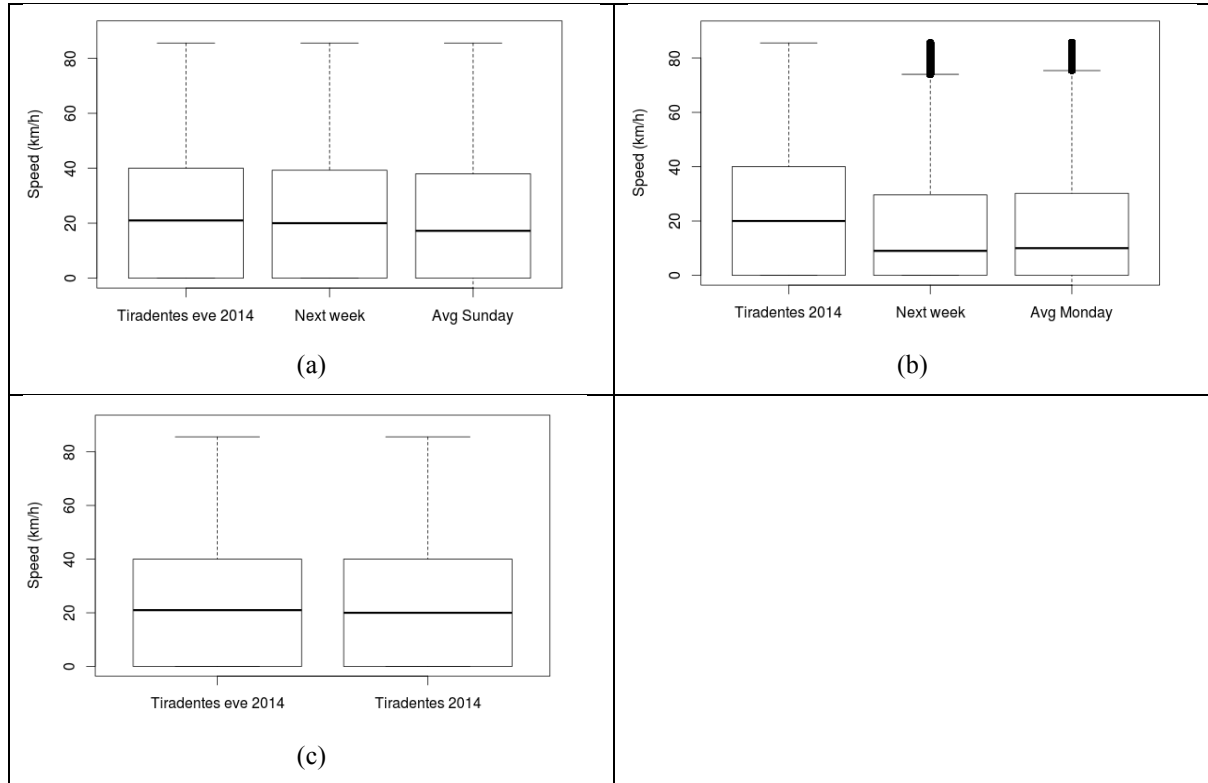


Figure 28. (a) Tiradentes eve 2014 speed comparison with next week Sunday and average Sunday. (b) Tiradentes 2014 speed comparison with next week Monday and average Monday. (c) Tiradentes eve and Tiradentes day speed comparison on 2014.

Figure 29 shows the position diff of the line 266 between the average Monday and Tiradentes 2014 holiday, generated by our position heat map diff tool. It is visible that the majority of the heat map is filled with red. This indicates that most of the line itinerary has less buses working on the average Monday than on Tiradentes 2014 holiday.

Figure 30 shows the speed diff of line 266 between the average Monday and Tiradentes 2014 holiday, generated by our speed heat map diff tool. It is visible that most of the heat map is filled with green. This indicates that the traffic is predominantly better on this line on the average Monday than on Tiradentes 2014 holiday. The unexpected result may have occurred because people traveled on the weekend to enjoy both the weekend and the holiday and got back home on the Monday, the day of Tiradentes holiday.

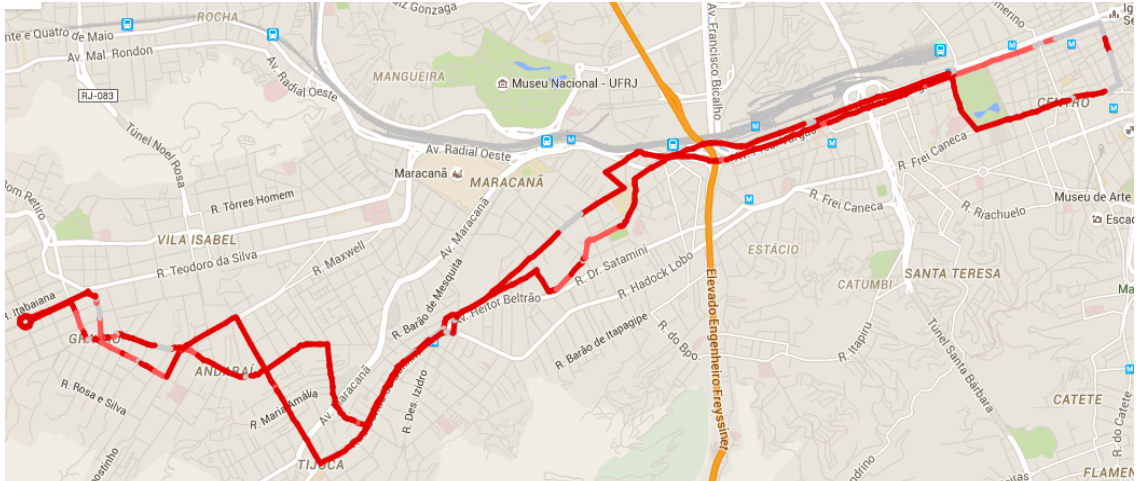


Figure 29. Position diff of line 266 between the average Monday and Tiradentes 2014 holiday

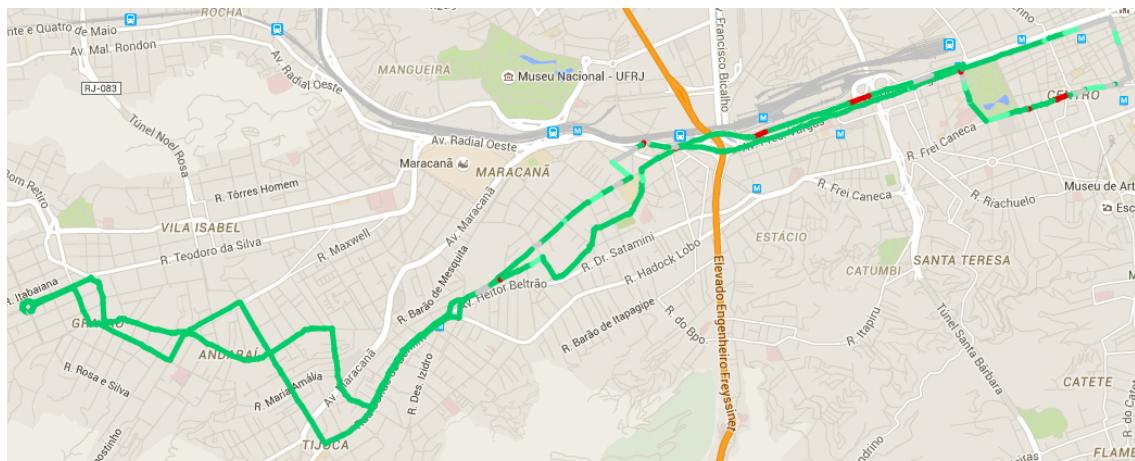


Figure 30. Speed diff of line 266 between the average Monday and Tiradentes 2014 holiday

On 2015 Tiradentes holiday did not behave the same as on 2014. Figure 31 (a) presents the speed comparison between Tiradentes eve (a Monday), previous week Monday, next week Monday and the average Monday. It is visible that the speed is slightly higher on Tiradentes eve than on the previous week Monday, the next week Monday and the average Monday. As on Tiradentes day on 2014, the holiday eve traffic is only a little better than the average of the equivalent day of week.

Figure 31 (b) presents the speed comparison between Tiradentes, previous week Tuesday, next week Tuesday and the average Tuesday. It is visible that the speed is much higher on Tiradentes day than on previous week Tuesday, next week Tuesday and the average Tuesday. As on Tiradentes day on 2014, the traffic is better on the holiday than on the average of the equivalent day of the week. Figure 31 (c) shows the speed comparison between Tiradentes eve and Tiradentes day on 2015. It is visible that the speed is much higher on Tiradentes day than on Tiradentes eve. In this case, the holiday eve behaves like an average equivalent day of week and the holiday day have a similar behavior to a weekend.

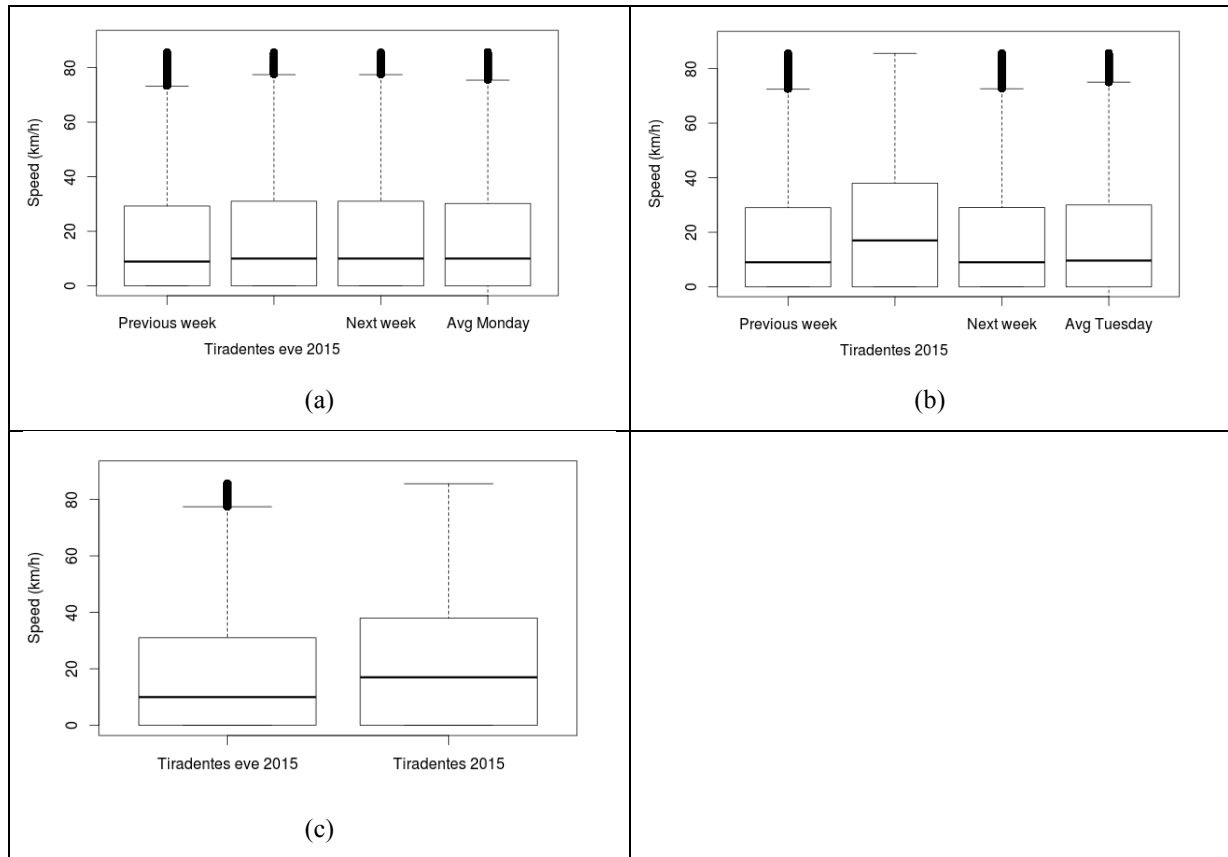


Figure 31. (a) Tiradentes eve 2015 comparison with previous week, next week and average Monday. (b) Tiradentes 2015 comparison with previous week, next week and average Tuesday. (c) Tiradentes eve and Tiradentes day speed comparison on 2015.

On 2014, the Labor Day holiday did not behave similarly to Tiradentes 2014 or Tiradentes 2015. Figure 32 (a) presents the speed comparison between Labor Day eve, previous week Wednesday, next week Wednesday and the average Wednesday. It is visible that the speed is much higher on the previous week Wednesday than on Labor Day eve. This fact happens because the previous Wednesday, April 23, was a holiday. We can also notice that the speed on Labor Day eve is near the same of the next week Wednesday and the average Wednesday. Different from what happened on Tiradentes eve on both years, the Labor Day eve traffic is very similar to the average of the equivalent day of week.

Figure 32 (b) presents the speed comparison between Labor Day, previous week Thursday, next week Thursday and the average Thursday. It is visible that the speed is very similar on all the compared dates. Different from what happened on Tiradentes on both years, the traffic on the holiday is similar to the average of the equivalent day of the week.

Figure 32 (c) shows the speed comparison between Labor Day eve and Labor Day on 2014. It is visible that the speeds on Labor Day and on Labor Day eve are very similar. In this case, the holiday eve and the holiday day seem to have similar behavior to a common weekday.

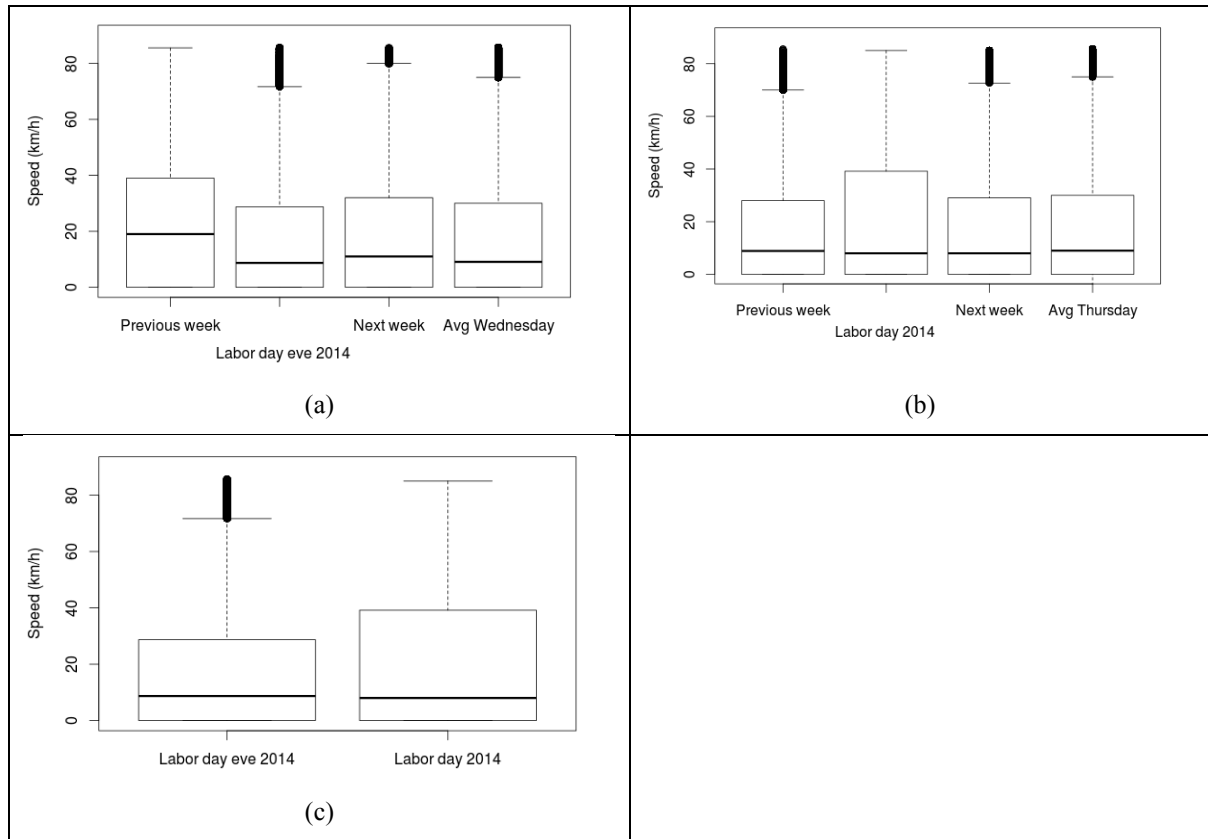


Figure 32. (a) Labor Day eve 2014 comparison with previous week, next week and average Wednesday. (b) Labor Day 2014 comparison with previous week, next week and average Thursday. (c) Labor Day eve and Labor Day speed comparison on 2014.

On 2015 Labor Day holiday did not behave the same as the other holidays. Figure 33 (a) presents the speed comparison between Labor Day eve, previous week Thursday, next week Thursday and the average Thursday. It is visible that the speed is much higher on the previous week Thursday than on Labor Day eve because of the same holiday on previous Thursday, April 23. We can also notice that the speed on Labor Day eve is slightly lower than the next week Thursday and the average Thursday.

Different from what happened on the previous year, the Labor Day eve traffic is slightly slower than the average of the equivalent day of week. Figure 33 (b) presents the speed comparison between Labor Day, previous week Friday, next week Friday and the average Friday. It is visible that the speed is much higher than on all the compared dates. As on Tiradentes day on both years, the traffic is better on the holiday than on the average of the equivalent day of week. Figure 33 (c) shows the speed comparison between Labor Day eve and Labor Day on 2015. It is visible that the speed is much higher on Labor Day than on Labor Day eve. In this case, the holiday eve and the holiday day seem to have similar behavior to a common weekday. In this case, the holiday eve behaves like an average equivalent day of week and the holiday day has similar behavior to a weekend.

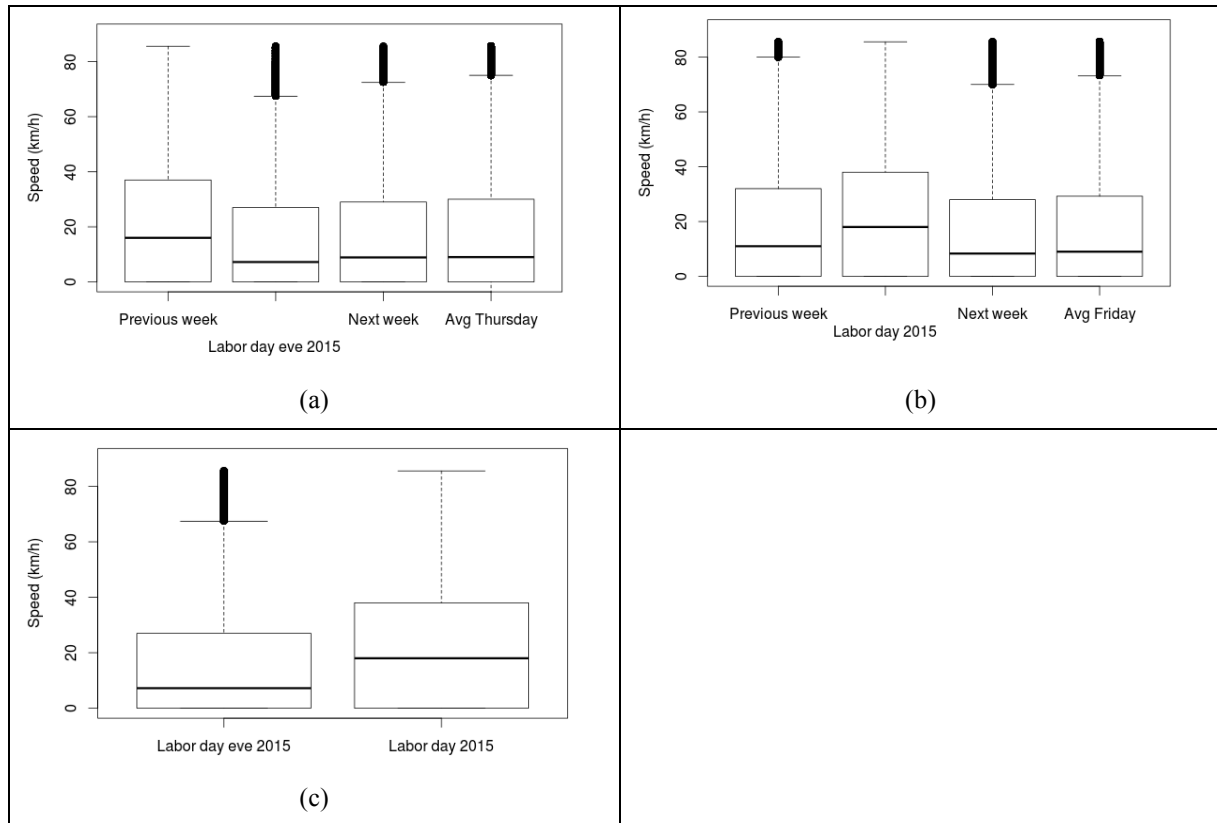


Figure 33. (a) Labor Day eve 2015 comparison with previous week, next week and average Thursday. (b) Labor Day 2015 comparison with previous week, next week and average Friday. (c) Labor Day eve and Labor Day speed comparison on 2015.

In all cases the traffic on holidays eves behave like the average of the equivalent day of week with a low variation. The traffic on the holiday is, in general, much better than the equivalent day of week and very similar to a weekend.

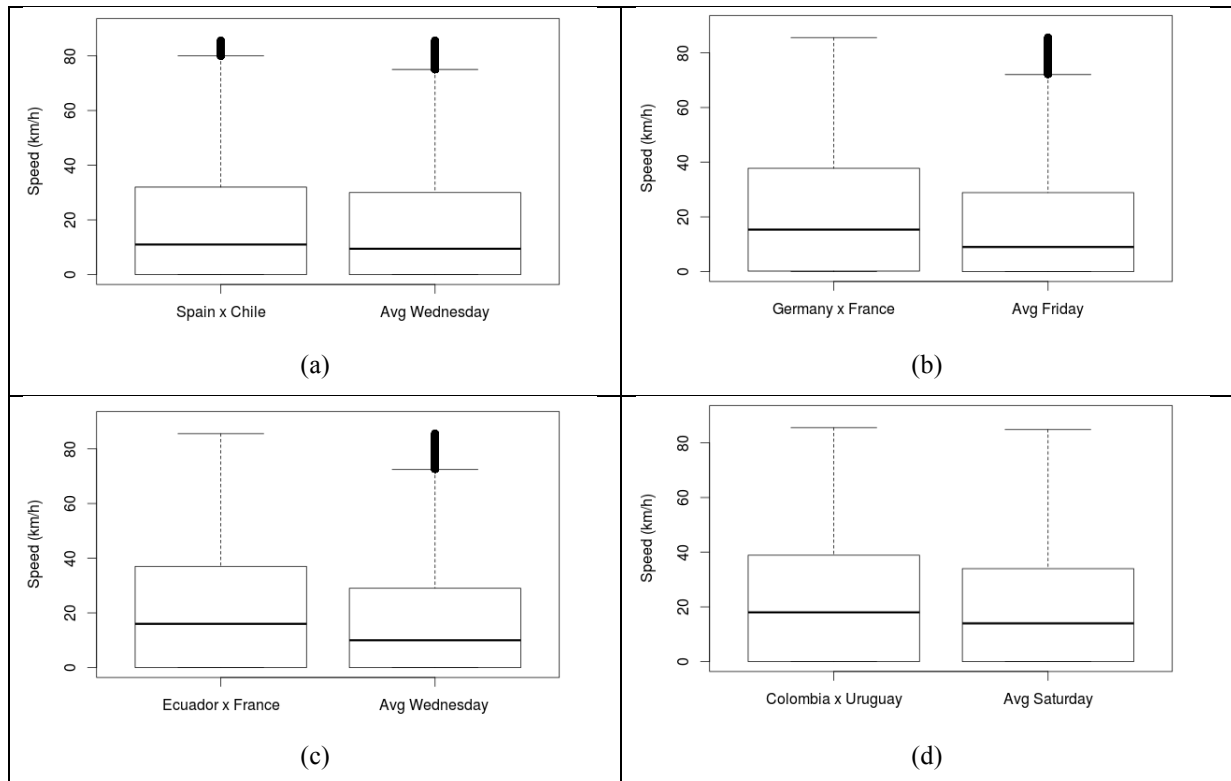
4.5 RQ5 - WHICH IS THE BEHAVIOR OF THE TRAFFIC DURING A LARGE EVENT IN THE CITY?

To answer this question we picked the days of six of the seven matches that happened at Maracanã Stadium on the 2014 Soccer World Cup. We do not have data on the day of the match Argentina x Bosnia and Herzegovina, which happened on June 15th because of some failure on our server or a failure on the Data Rio server. The other matches were: Spain x Chile on the Wednesday, June 18th; Belgium x Russia on Sunday, June 22nd; Ecuador x France on the Wednesday, June 25th; Colombia x Uruguay on the Saturday, June 28th; Germany x France on the Friday, July 4th; and Germany x Argentina, the World Cup final, which happened on the Sunday, July 13th. It is important to notice that, to avoid traffic problems, these days were declared holiday by the mayor of Rio de Janeiro city (G1 RIO, 2014).

Figure 34 (a), (e) and (f) shows respectively, the Spain x Chile match day speed comparison with the average Wednesday, the Belgium x Russia match day comparison with the average Sunday and the World Cup Final match day comparison with the average Sunday. On the days of these matches the speed was a little higher than the average equivalent day of week.

Figure 34 (b), (c) and (d) shows respectively, the Germany x France match day comparison with the average Friday, the Ecuador x France match day comparison with the average Wednesday and the Colombia x Uruguay match day comparison with the average Saturday. On the days of these matches the speed was considerably higher than the average equivalent day of week.

In all cases the traffic on the days of matches of the World Cup were better than the equivalent day of week. The cases with higher discrepancy were the matches happened on weekdays. As the match days were holidays, the traffic behaves very similar to the holidays analyzed in the previous section.



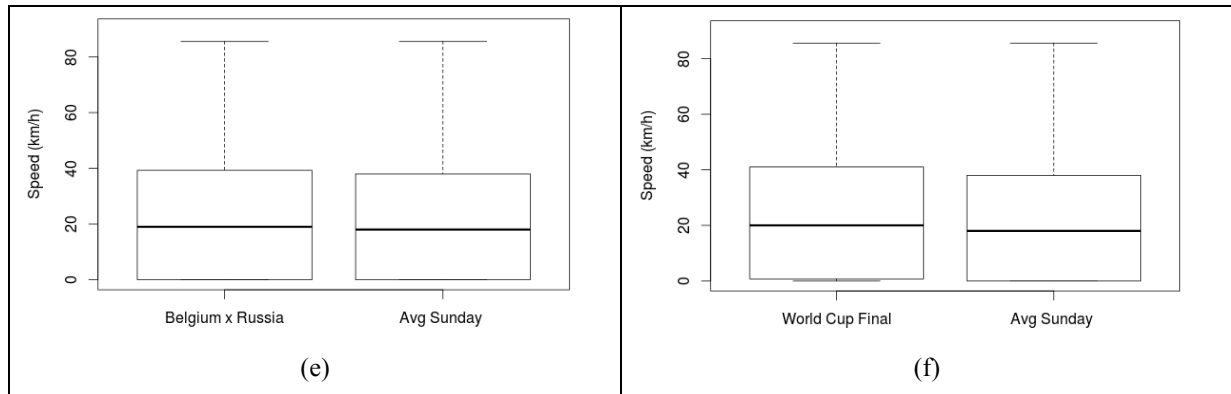


Figure 34. (a) Spain x Chile match day comparison with the average Wednesday. (b) Germany x France match day comparison with the average Friday. (c) Ecuador x France match day comparison with the average Wednesday. (d) Colombia x Uruguay match day comparison with the average Saturday. (e) Belgium x Russia match day comparison with the average Sunday. (f) World Cup Final match day comparison with the average Sunday.

4.6 RQ6 - HOW DOES THE TRAFFIC CONDITIONS VARY IN DIFFERENT MONTHS?

To answering this question, we picked all the data in our database. The collected data were the 19 months from April 2014 until October 2015. Figure 35 shows the speed comparison between the months of the year. January has the best traffic of the year and December has the worst traffic of the year.

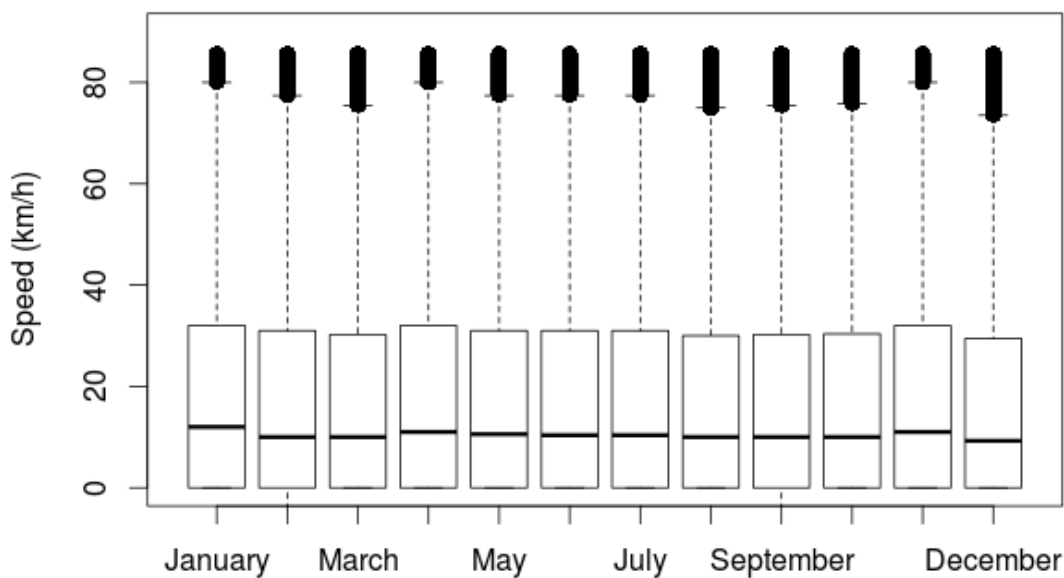


Figure 35. Speed comparison between the months of the year

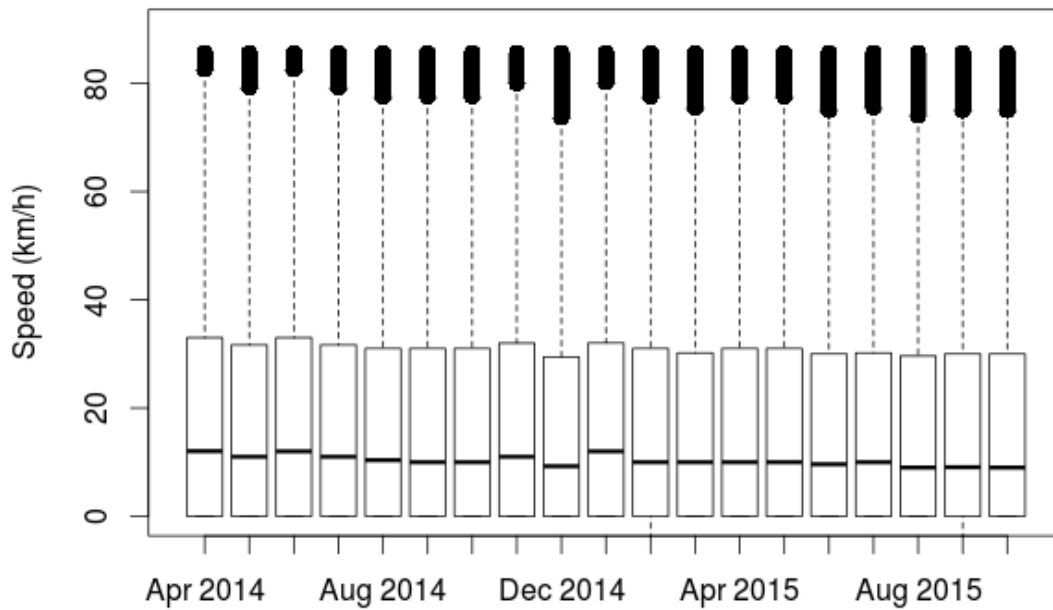


Figure 36. Speed comparison between each of the collected months

Figure 36 shows the speed comparison between each of the collected months. The best traffic of our database is on January 2015 and the worst is on December 2014. Despite January 2015 was the best traffic of our entire database, the months of 2015 present worse traffic than the equivalent months of 2014.

The months of January, February and March were only collected on 2015 and the months of November and December were only collected on 2014. Even though we loaded more than one billion records, it is hard to identify patterns of the traffic through the year because we only had 19 months to compare.

We did not have enough data to identify patterns of the traffic through the year because we only had 19 months to compare.

4.7 FINAL REMARKS

In this chapter we have presented the results of the study made using the approaches and tools shown in the previous chapter to answer the research questions presented in the introduction of this work. In the next chapter we present the conclusion of this work.

CHAPTER 5 – CONCLUSION

5.1 CONTRIBUTIONS

This work contributes by providing tools to efficiently load and clean bus data, distinguishing abnormal data and valid data. Along with the visual analysis tools, this work uses a set of 1,253,724,203 records to allow us answering the research questions presented in the introduction and helping us to understand the bus traffic in Rio de Janeiro.

Using these tools, we contributed by identifying several kinds of abnormalities on the bus data provided by the Rio de Janeiro city hall. Knowing that only 32% of the data was considered correct makes us observe that the data needs to be submitted to several cleaning methods before it can be used with thrust in applications.

Moreover we could find patterns in the traffic data in the days of the week. We identified that the weekdays have similar behaviors, with a low variation between them. The weekends have better traffic compared to the weekdays. Saturday presents a considerable increase in the average speeds compared to the weekdays and Sunday presents a relevant increase compared to Saturday.

Further we found patterns in the traffic data looking through the hours of the day. The traffic presents high variation through the day. The range with lower speed is the end of the afternoon hours and beginning of the night hours. The range with higher speed is the early morning hours.

Another analysis was comparing holidays with the average of the common days. We used two holidays and it's eves in this analysis: Tiradentes and Labor Day in 2014 and 2015. We have observed that in all cases the traffic on holidays eves behave like the average of the equivalent day of week with a low variation. The traffic on the holiday is, in general, much better than the equivalent day of week and very similar to a weekend.

The next study was the traffic during a large event in the city. We analyzed the traffic in the days of the matches happened at the Maracanã stadium for the soccer World Cup in 2014. We notice that in all cases the traffic on the days of matches of the World Cup was better than the equivalent day of week. The cases with higher discrepancy were the matches that happened on weekdays. As the match days were holidays, the traffic behaves very similar to the holidays analyzed in the previous section.

The last proposed analysis was comparing the months of the year. We analyzed all the data loaded over the proposed period. In this topic no visible patterns were found. We came to

the conclusion that we did not have enough data to identify patterns of the traffic through the year because we only had 19 months to compare.

5.2 THREATS TO VALIDITY

Despite the effort made to provide a consistent analysis, we have identified some threats to validity. The threats found were two: lack of data and cleaning methods not identified.

Even though we have used a large amount of data (1,253,724,203 records), our last study, comparing the traffic through the months of the year, clearly suffered from lack of data. Because of this, we could not get any results from this analysis. The same way, it is possible that we came to misleading conclusions because we thought that we had enough data for some other question in the study when in fact we did not. This may have happened on RQ4 when we study only four holidays and holidays' eves.

The second threat to validity of our study is the possibility of us needing some cleaning process that was not done. As presented earlier, only 32% of the data collected from our source was considered valid to this study. Despite all of the cleaning methods used in this work, it is possible that we did not identified some other cleaning needed and then we would have had even less valid data, but this data would have been more reliable.

5.3 FUTURE WORK

For future work, we plan to automatically identify where are the garages and extend the loading system. Once having this feature, it will be possible to solve the false negative cases of the *Bus out of line itinerary* filtering method when the garage is located inside the bounding box of the bus line.

Another possible future work is the identification of the line that the bus is working when the JSON file does not provide this information. This feature could bring a wide increase of the valid data since 29% of the records were discarded in our experiment by not having the bus line information filled in the JSON files.

The UI module can be extended and include a more user-friendly UI for the loading module. This would make easier to choose the files to be processed and would let the progress of the loading clearer.

Another work on the UI is including interfaces for showing and exporting the charts generated by R. In this study we used R Studio to generate the boxplots and the correlation

chart presented in Chapter 4. It would be easier to the users if they could generate these charts inside the same UI module they use to generate the heat maps.

The database can be changed from PostgreSQL to a more appropriate tool. Given the large amount of records, using a NoSQL database tool could result in great improvements in the performance of loading and querying data.

In a further work, it would be useful to analyze a larger amount of data. Analyzing more data could allow finding new patterns over the months and over the years, which was not possible with the 19 months analyzed in this work.

Finally, another extension could be using new sources of data to increase the analysis possibilities. Adding new georeferenced data sources as road works and climate events would allow new studies and a better knowledge of the traffic.

REFERENCES

- ABNT. **NBR 14724:2005 - Informação e documentação - Trabalhos acadêmicos - Apresentação**. Rio de Janeiro: Associação Brasileira de Normas Técnicas, 2005.
- ALBUQUERQUE, F. DA C. et al. A proactive application to monitor truck fleets. **2013 IEEE 14th International Conference on Mobile Data Management**, v. 1, p. 301–304, 2013.
- ANDRADE, L. S.; CRUZ, S. M. S. DA. *BusInRio: Explorando Dados Abertos de Transporte Público do Município do Rio de Janeiro*. 2015.
- ANDRIENKO, G.; ANDRIENKO, N.; WROBEL, S. Visual Analytics Tools for Analysis of Movement Data. **ACM SIGKDD Explorations Newsletter**, v. 9, n. 2, p. 38–46, 2007.
- BARBOSA, L. et al. Vistradas: Visual Analytics for Urban Trajectory Data. **GeoInfo**, p. 174–179, 2014.
- BESSA, A. et al. RioBusData: Outlier Detection in Bus Routes of Rio de Janeiro. **arXiv preprint arXiv:1601.06128**, 2016.
- BRADFORD W. PARKINSON; JAMES J. SPILKER. **Global Positioning System: Theory and Applications**. [s.l.: s.n.]. v. 1
- DONNA J. PEUQUET. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. **Annals of the Association of American Geographers**, v. 84, n. 3, p. 441–461, 1994.
- FERREIRA, N. et al. Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 2149–2158, dez. 2013.
- G1 RIO. **Prefeitura decreta feriado em dias úteis com jogos da Copa no Rio**. Disponível em: <<http://g1.globo.com/rio-de-janeiro/noticia/2014/03/prefeitura-decreta-feriado-em-dias-uteis-com-jogos-da-copa-no-rio.html>>. Acesso em: 3 maio. 2016.
- GEMS. **Zip files containing Rio de Janeiro City bus fleet data**. Disponível em: <<http://sel.ic.uff.br/bus/>>. Acesso em: 7 jul. 2016.
- GENNADY ANDRIENKO; NATALIA ANDRIENKO. Spatio-temporal aggregation for visual analysis of movements. **Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on**, p. 51–58, 2008.
- JALALI, S.; WOHLIN, C. Systematic literature studies: database searches vs. backward snowballing. **6th ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM**, 2012.
- JULIANA FREIRE et al. Riding from Urban Data to Insight Using New York City Taxis. 2014.

LÉ CUÉ, F. et al. Semantic traffic diagnosis with Star-City: Architecture and lessons learned from deployment in Dublin, Bologna, Miami and Rio. **International Semantic Web Conference**, p. 292–307, 2014.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LIU, S. et al. **A visual analytics system for metropolitan transportation**. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. **Anais...ACM**, 2011

LIU, S. et al. VAIT: A Visual Analytics System for Metropolitan Transportation. **IEEE Transactions on Intelligent Transportation Systems**, 2013.

LU, C.-T.; BOEDIHARDJO, A. P.; ZHENG, J. **AITVS: Advanced Interactive Traffic Visualization System**. Proceedings of the 22nd International Conference on Data Engineering, 2006. ICDE '06. **Anais...** In: PROCEEDINGS OF THE 22ND INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 2006. ICDE '06. abr. 2006

MARCOS R. VIEIRA et al. USapiens: A System for Urban Trajectory Data Analytics. **2015 16th IEEE International Conference on Mobile Data Management**, v. 1, p. 255–262, 2015.

PRJ. **data.rio - Portal de dados abertos da Prefeitura do Rio**. Disponível em: <<http://data.rio.rj.gov.br/>>. Acesso em: 1 jul. 2016.

PU, J. et al. **T-Watcher: A New Visual Analytic System for Effective Traffic Surveillance**. Mobile Data Management (MDM), 2013 IEEE 14th International Conference on. **Anais...IEEE**, 2013

SHEKHAR, S. et al. Cubeview: a system for traffic data visualization. **Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on**, p. 674–678, 2002.

THIAGARAJAN, A. et al. Cooperative Transit Tracking using Smart-phones. **Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems**, p. 85–98, 2010.

YAN, Z. et al. **Automatic Construction and Multi-level Visualization of Semantic Trajectories**. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. **Anais...**: GIS '10. New York, NY, USA: ACM, 2010. Disponível em: <<http://doi.acm.org/10.1145/1869790.1869879>>. Acesso em: 26 nov. 2014

ZICHENG LIAO; YIZHOU YU; BAOQUAN CHEN. Anomaly detection in gps data based on visual analytics. **Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on**, p. 51–58, 2010.