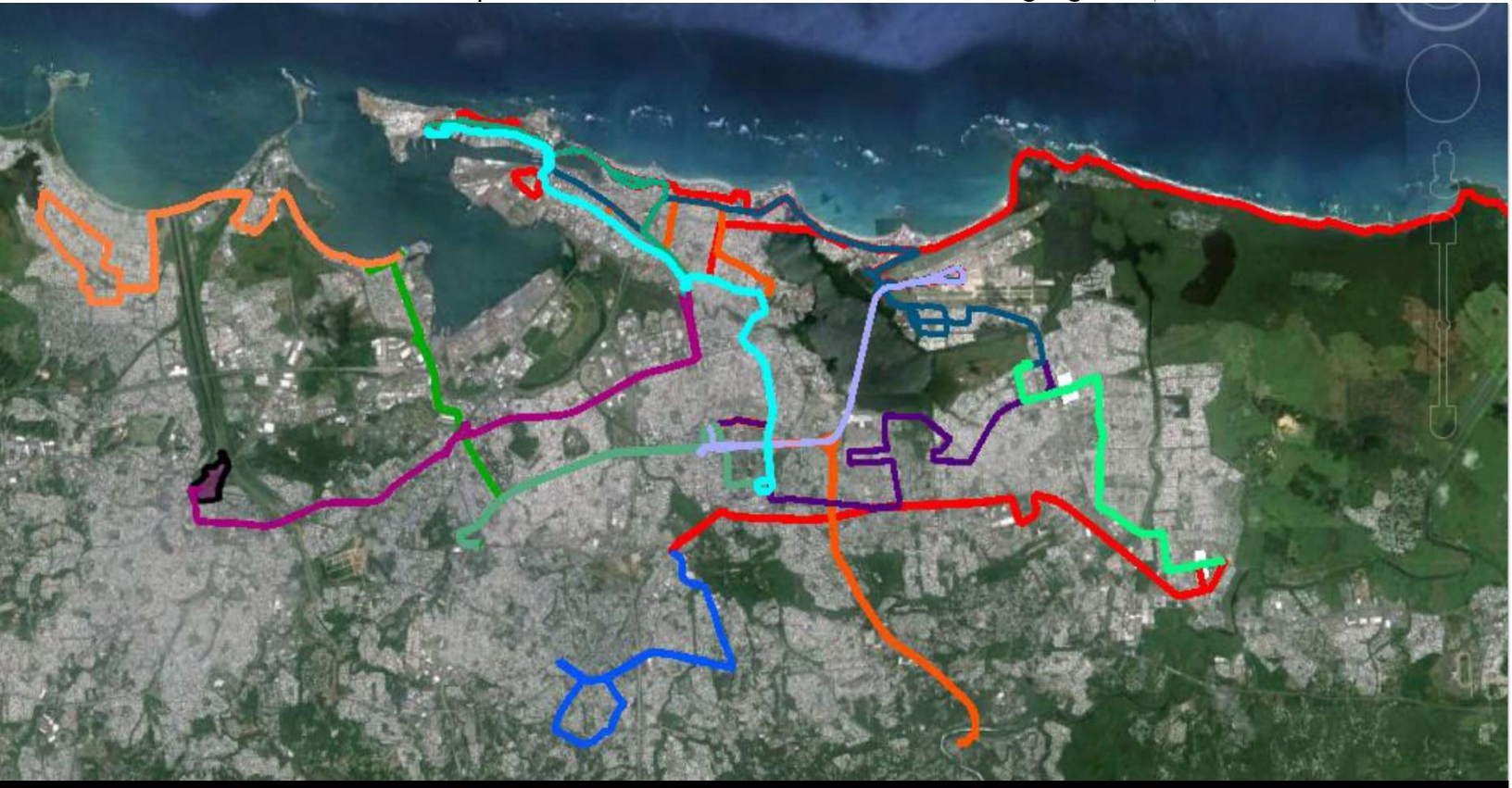
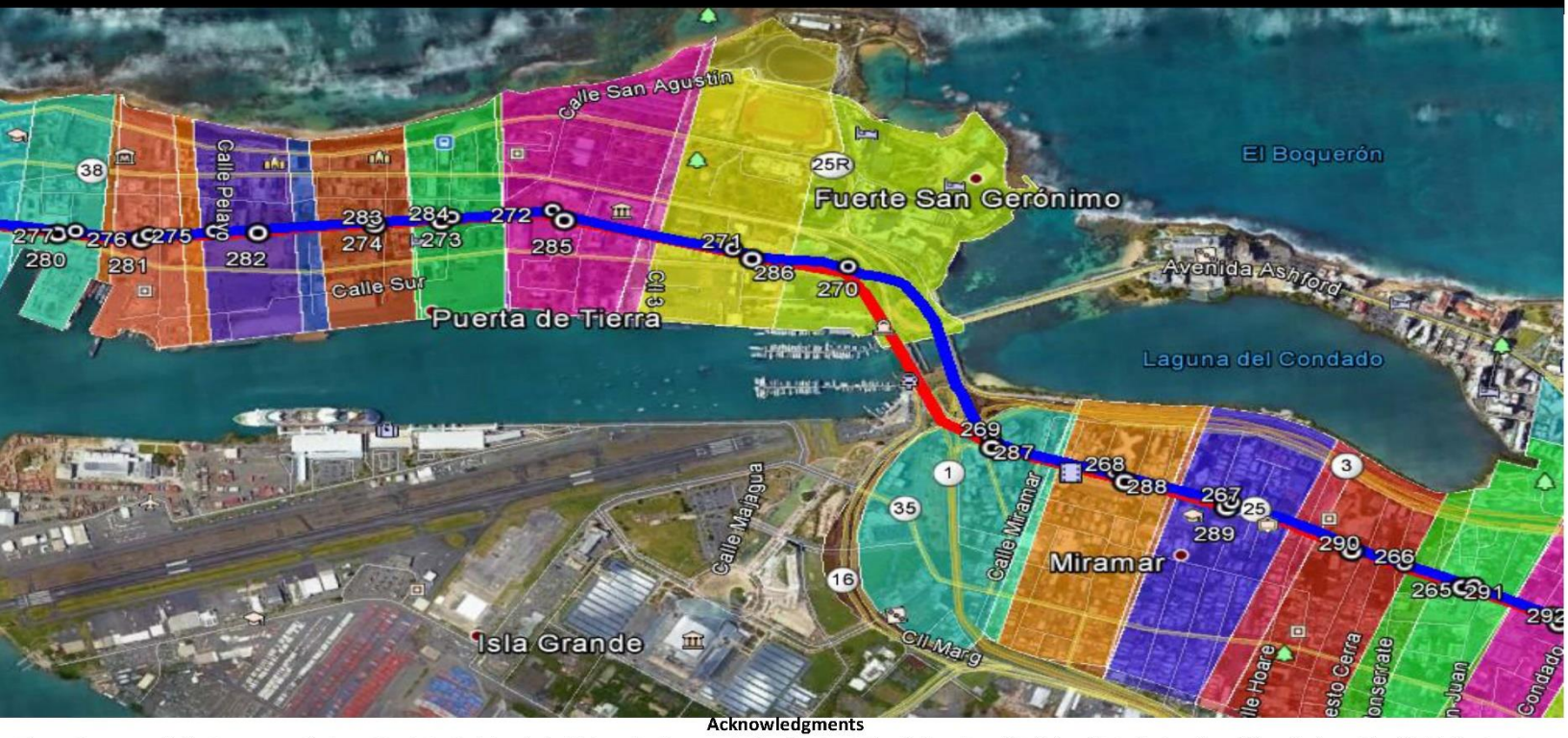


DEVELOPMENT OF TRANSIT PERFORMANCE MEASURES USING BIG DATA

Didier Valdes Ph.D.: Professor, Department of Civil Engineering and Surveying, University of Puerto Rico-Mayagüez
Ivette Cruzado, Ph.D.: Associate Professor, Department of Civil Engineering and Surveying, University of Puerto Rico-Mayagüez
Juan Martinez: Ph.D. Candidate, Department of Civil Engineering and Surveying, University of Puerto Rico-Mayagüez: juan.martinez@upr.edu
Yindhira Taveras: Ph.D. Candidate, Department of Civil Engineering and Surveying, University of Puerto Rico-Mayagüez: Yindhira.taveras@upr.edu



Acknowledgments

The authors would like to express their gratitude to the TransInfo University Transportation Center at the University of Buffalo – State University of New York and the Civil Infrastructure Center at the University of Puerto Rico a Mayaguez for their financial support.

TransINFO
University Transportation Center

DEVELOPMENT OF TRANSIT PERFORMANCE MEASURES USING BIG DATA

Report date: June, 2017

Didier, Valdes, PhD, Professor, Civil Engineering and Surveying Department, UPRM
Ivette, Cruzado, PhD, Assistant Professor, Civil Engineering and Surveying Department, UPRM
Juan, Martínez, PhD Civil Engineer Graduate Student, Civil Engineering and Surveying Department, UPRM
Yindhira, Taveras, PhD Civil Engineer Graduate Student, Civil Engineering and Surveying Department, UPRM

Prepared by:
Civil Engineering and Surveying Department
Calle Post, Mayagüez Puerto Rico
University of Puerto Rico-Mayagüez
PR-108, Mayagüez, 00682, Puerto Rico

Prepared for:
Transportation Informatics Tier I University Transportation Center
204 Ketter Hall
University at Buffalo
Buffalo, NY 14260

Report No.	Government Accession No.	Recipient's Catalog No.

4. Title and Subtitle DEVELOPMENT OF TRANSIT PERFORMANCE MEASURES USING BIG DATA	**5. Report Date** June 2017
	6. Performing Organization Code
7. Author(s) Didier Valdés Ivette Cruzado Juan Martínez Yindhira Taveras	**8. Performing Organization Report No.**
9. Performing Organization Name and Address Civil Engineering and Surveying Department Calle Post, Mayagüez Puerto Rico University of Puerto Rico-Mayagüez PR-108, Mayagüez, 00682, Puerto Rico	**10. Work Unit No. (TRAIS)**
	11. Contract or Grant No. DTRT13-G-UTC48
12. Sponsoring Agency Name and Address US Department of Transportation Office of the UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590	**13. Type of Report and Period Covered** **Project start date: January, 2014** **Project end date: December, 2017**
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This project presents the development of real-time performance measures obtained by using Big Data generated by AVL/GPS systems installed in public transit vehicles merged with transportation demand related data available from other sources, including the Census. This type of synthesized, real-time information can improve decision-making at the operational and planning levels. Using the developed performance measures, the system can be evaluated in real-time during day-to-day operations, as well as in short and medium-term planning. This project was carried out using the data from the Puerto Rico Metropolitan Bus Authority system (AMA, for its acronyms in Spanish), which is the main public transit operator in the San Juan Metropolitan Area of Puerto Rico.

17. Key Words	18. Distribution Statement
Big Data, Performance Measures, Transit	No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	103	

ABSTRACT

All transportation systems can be evaluated using metrics that demonstrate their performance. The estimation of these metrics is based on data which could be obtained through different methods or equipment, such as GPS. The equipment records updated information about vehicle units' time and location.

The fundamental purpose of this project is to propose a methodology to obtain real-time data from large databases, generated by GPS-based Automatic Vehicle Location AVL systems installed in public transit vehicles, and merge that information with transportation demand related data available from other sources, including the Census, to estimate system performance measures and propose new metrics to help improve service with the use of "Big Data" management programs, such as Knime-Analytics.

This type of synthesized, real-time information can improve decision making at the operational and planning levels. The performance measures presented in this project report can be used to evaluate the system in real-time day to day operations, as well as in short and medium-term planning.

This project is taking place using the Metropolitan Bus Authority system (whose Spanish acronym is AMA), which is the main public transit operator in the San Juan Metropolitan Area of Puerto Rico. Performance metrics are calculated for AMA Route 5. The service level obtained using the proposed measures correlate with the boarding levels observed in the field. One recommendation is to expand this type of study to the entire system.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Graphics

1 INTRODUCTION

With recent advances in technology, we have entered the Big Data world, where it is possible to acquire a great amount of information at a relatively low cost or effort. Several transit systems have Automatic Vehicle Location (AVL), Automatic Fare Collection (AFC), and Automatic Passenger Counters (APC) in all or a portion of their vehicles. With the development of Global Positioning Systems (GPS), it is possible to include the location of the activity centers and their accessibility. Furthermore, Google Transit can monitor the mobility between these centers of activity in real-time. Recent research has discussed the possibility of using cellular location data to analyze the mobility of a population (Widhalm *et al.*, 2015).

These are just some examples of how recent technological developments can be used to improve transit system analysis, operation, and design. The information gathered can be useful as archived data (for planning and long-term analysis) and in real-time (for daily operation decisions, and instant quality analysis). In the past, the industry heavily relied on the use of relatively small samples; from which findings were extrapolated for the whole phenomenon. This approach meant a great deal of uncertainty and prevented real-time application. However, with current technology, it is possible to measure the performance of a transit system and the daily fluctuations of its surrounding activity centers; not only with sample data, but with more comprehensive information as well.

This research aims to expand scientific knowledge in the field of transportation engineering, particularly with regards to performance measures using real-time data. Transit Performance Measures have sparked interest in recent years. Due to rapid technological advancement, new ways of measuring performance have also been discussed. There is ample research on performance metrics, where researchers and operators have developed over 400 performance metrics. The application of AVL-APC archived data has also been widely researched. The development of a data-processing framework to process a massive amount of transit data, including vehicle location, passenger count, and electronic fare transactions, has been accomplished by researchers and operators. In 2008, the New York Metropolitan Transit Authority (MTA) took the first steps in making its transit information more accessible to the public. The MTA currently presents On-time performance and Mean Distance between Failure and Ridership.

Performance measurement and peer comparison are valuable tools to efficient and proactive management. The transit industry has relied on limited, general, and aggregated measures for reporting its performance to regulatory agencies. However, with the development of new technologies, it is now possible to measure the performance of a transit system considering data from the whole system, as opposed to just sample data.

This report presents the results and information obtained at the San Juan Metropolitan Area (SJMA). In the present, AMA complies with FTA mandates in terms of collecting data on system performance but without the previously discussed technological tools. AMA recently bought AVL and APC devices to install on their vehicles, but they lack the computational tools needed to take advantage of them. Our main goal is to present AMA with a practical and feasible option to redesign their information gathering system. Said analyses target increased passenger demand by modifying the activities system and the operational parameters, or performance measures.

1.1 PROBLEM JUSTIFICATION

The San Juan Metropolitan Area (SJMA) is an urbanized area surrounding Puerto Rico's main city and capital: San Juan. It includes over ten (10) municipalities and a population larger than 1 million. The Metropolitan Bus Authority (AMA) is the governmental transit agency serving the SJMA. In the present, AMA complies with FTA mandates concerning system performance data collection but without using the previously mentioned technological tools. AMA bought AVL and APC devices to install on its vehicles, but lacks the computational tools needed to take advantage of them.

With their current system, real-time GPS data was available for the buses, so the exact locations of all the buses in service were registered in 10-second intervals. Analyses continue being developed with the recently acquired data. The results can be helpful to identify strategies for improving the AMA system's service, which is described below

Description of AMA System	➢ AMA system has 23 fixed-routes with stationary stops over the SJMA. ➢ AMA has identified 15 bus routes with the highest ridership. ➢ The vehicles serving these routes, with a capacity of 70 people each, are categorized as standard high-floor models with steps at the door that sometimes imply longer dwell times. ➢ The operational environment is mostly mixed traffic, but exclusive lanes are used in some segments of highly congested streets. ➢ The network covers over 100 miles (See Figure 2), with an estimated ridership of 30,000 passengers per day.

Figure 1 Description of AMA System

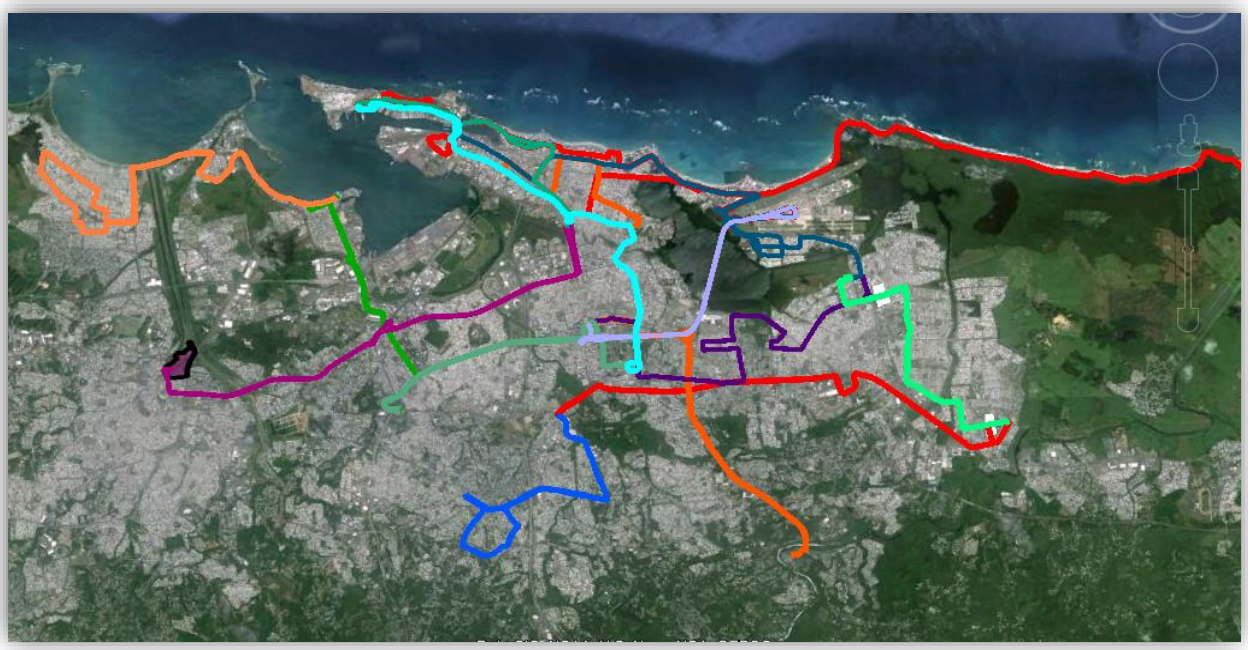


Figure 2 Map of AMA System Routes

This study aims to develop new practices for performance measurement assessment by drawing from previously developed knowledge in the field. In order to address performance issues, this study uses the perspective of different engineering fields, such as computer and electrical engineering; thus, giving a lead in the technological advances in the transportation engineering field.

The implications of this study could range from maintaining and enhancing physical systems of transportation by using advanced technology, to developing new practices pertaining to mobility and access for all users by integrating a social justice perspective. Both inferences have implications for policy and theory development within the context of transportation systems' efficiency and cost effectiveness.

This study also aims to cultivate leadership skills by integrating students in the decision-making processes needed to develop new performance measurements in the transportation engineering field.

1.2 MAIN GOAL AND SPECIFIC OBJECTIVES

The objective of this research is to develop performance measures for transit systems based on big data obtained through GPS/AVL systems combined with activity systems data gathered from various sources. This study uses data from the Puerto Rico Metropolitan Bus System (AMA) combined with publicly available data from the activity system along the studied routes. Since real-

time GPS data was available for the buses, we could know knowing the exact locations of all the buses in service as they were registered every 10 seconds. Also, a passenger boarding study was conducted and information about activity centers was digitalized.

The specific objectives include:

1) Displaying buses' GPS location data to understand how the system operated.
2) Analyzing the archived data to prepare performance reports.
3) Calculating a specific real-time performance metric that may be used to determine real-time schedule adjustments based on the field's actual conditions.
4) Determine if there is a direct correlation between activity centers and passenger boarding that can be used in conjunction with the operational data to improve the system's efficiency.

Various analyses have been developed with the big data acquired for this study. The results can be helpful in identifying strategies for improving the AMA system's service and increasing ridership boarding.

1.3 REPORT CONTENTS AND ORGANIZATION

This report explains the work performed for the duration of the research and is divided into the following sections: Introduction, Literature Review, Methodology, Data, Performance Measures (Current, Proposed and Activity System Integration), Conclusions, References, and Appendix.

2 LITERATURE REVIEW

In this research, the first step consisted of a literature review about developing performance measures for public transportation using big data. Past trends in the research and development of new performance measures were also studied and are included in this section of the report.

Transit Systems play a vital role in terms of the economy, energy, and the environment, but they are also very important in terms of social equity, mobility, and access to jobs, education, and services, among others. After all, one of transit's major roles is to provide essential mobility for those who are too young, too old, or otherwise unable to drive due to physical, mental, or financial disadvantages (Transportation Research Board., 2013). Offering the best possible transit service is not only an engineering problem, but a social justice one. Transit systems operators need to monitor the system's performance to assess its quality and improve it when necessary.

Transit Performance Measures is a topic that has received considerable attention in recent years, although it was frequently discussed before the first publication of the *Transit Capacity and Quality of Service Manual (TCQSM)* in 1999 (Kittelson & Associates, et al., 1999). Since technology has advanced rapidly, new ways of measuring performance have also been discussed. In the last 15 years, the TCRP project published two additional editions of the manual first in 2003 and later in 2011. The *TCRP Report 88* provides useful information on more than 400 transit performance measures.

Shannon and Bellisio (2013) indicate that performance measures encompass the collection, evaluation, and reporting of data that reveal how well an organization performs its functions and meets its goals and objectives. The measures included in this process should relate to the outcomes achieved by the organization. Also, performance metrics can be used to detect reasons for changes needed to fulfill the objectives of an organization. Therefore, performance metrics are used "to evaluate performance, identify opportunities for improvement, establish performance goals, and help guide expenditures and investments" (Shannon & Bellisio, 2013). They are instruments to help organizations better understand themselves in comparison to other, similar organizations. They are also used to weigh strengths and weaknesses, and based on this comparison, develop strategies for changes in services, planning and evaluation practices.

The TCRP Report 141 (Transportation Research Board, 2010) establishes that the importance of the performance metrics rests, not on their absolute values, but in its comparison with something else —for example, one's past performance, one's targeted performance, or comparable organizations' performance—to provide the context of "performance is good," "performance needs improvement," "performance is getting better," and so on.

The TCRP Report 165 (Transportation Research Board., 2013) includes the state-of-the-art standard in performance measures from users' point of view. This report explains the differences between four possible points of view: (a) user, (b) agency, (c) auto drivers, and (d) community. This document gathered different metrics in two broad categories: (1) availability, and (2) comfort and convenience. The TCRP Report 141 (Transportation Research Board, 2010) and the Highway Capacity Manual are the state-of-the-art standards in performance measures for private vehicle drivers who share the road with transit vehicles. A manual from the community's perspective doesn't presently exist, but researchers in this area have focused on developing metrics related to pollution mitigation, noise contamination reduction, and accessibility to labor or health center, etc.

Although the agency's perspective is possibly the most studied, it doesn't have a manual setting the standards for its performance. In its place, every operator has developed a system for monitoring performance. The TCRP Report 88 (Transportation Research Board., 2003) provided a list of over 400 metrics suitable for various systems. The closest set of standards for monitoring performance from the agency's perspective is a set of metrics that the FTA (Federal Transit Administration, U.S Department of Transportation., 2015) requires every agency to report on in the National Transit Database, which includes demographic information of the population, revenue hours, revenue miles, unlinked passenger trips, asset information, etc. (Transportation Research Board, 2003)

In general terms, this information may be useful to describe a system, but it doesn't always help to identify specific areas in need of improvement.

The use of Automatic Vehicle Location (AVL) and Automatic Passenger Counters (APC) have been researched in more depth in the last two decades. AVL-APC systems can gather an enormous quantity and variety of operational, spatial, and temporal data that — if captured, archived, and analyzed properly — holds substantial promise for improving transit performance by supporting improved management practices in areas such as service planning, scheduling, and service quality monitoring. The report offered guidance on five subjects:

1. Analyses that can use AVL-APC data to improve management and performance
2. AVL-APC system design that facilitates the capture of data with the accuracy and detail needed for off-line data analysis
3. Data structures and analysis software for facilitating analysis of AVL-APC data
4. Screening, parsing, and balancing automatic passenger counts
5. Use of APC systems for estimating passenger-miles for National Transit Database (NTD) reporting (Furth, et al. 2006).

Since information gathering has been increasing over time, the capacity of processing said data had to improve as well. Liao and Liu (2010) developed a methodological data-processing framework to process a massive amount of transit data, including vehicle location, passenger count, and

electronic fare transactions. The developed data analysis methodology can have several applications, including transit route performance measurement, to support decision making for transit planning and operation. Also, Saavedra, et al (2011) discuss the problem of unreliable data and develop a methodology that can assure its quality.

While Cevallos and Wang (2008) developed an SQL based data archiving and mining system to help improve operational efficiencies and quality of service, Muller & Furth (2001) developed a trip time analyzer to analyze the running time by link, delay by segment, and schedule and headway deviation by time point (TP). Others have studied the applicability of AVL and APC data to the estimation of travel time (Salek, et al. 2011). The study begins by developing a regression model that generates accurate estimates of average travel time for individual route segments based on segment length, number of intersections, and passenger activity in that segment.

Furthermore, Hickey, et al. (2014) explored advanced train control system data and vehicle location as informational tools to examine the potential reassigning and interchanging of light rail and three rail vehicles. Ji and Zhang (2013) proposed a strategy that monitors bus locations in real-time and estimates the time gaps between consecutive buses at a desired frequency Tribone, et al. (2014) assessed how the agency measured performance and then redesigned and advanced the agency's daily performance reports.

Portland, Oregon became the first city visualized on Google Transit in November of 2006 (Crout, 2011) which integrates performance measures data from a variety of sources including AVL and APC.

In 2008, the Metropolitan Transit Authority (MTA) of New York took the first steps in making its transit information more accessible to the public. Currently, the MTA presents on-time performance (OTP) statistics (the percentage of trains that were on time in a given month or year), Mean Distance Between Failure (MDBF) (how far trains and buses travel before breaking down), and ridership (the number of people choosing transit instead of another form of transportation) (Shannon & Bellisio, 2013).

In both examples, data analysis is made offline, not in real-time. Such analyses have proved to be extremely useful for planning purposes but not for real-time decision making by the operator agency.

The topics of performance metrics and visualization of GPS/AVL and big data in public transportation have been studied by other authors. A list of the additional literature considered in this project includes the work of Steward, et al. (2015) who presented the benefits of visualizing transit data Shi, et al. (2015) who identified transit passenger characteristics and travel time reliability using Automatic Fare Collection (AFC) data, Van Oort & Cats (2015), who presented an implementation of big data analysis to improve not only operations but also planning and

decision making, and Lock & Erhardt (2015) who presented the use of large, automatically connected data sources to measure and monitor transit system performance. Also, Czech & Turner (2014) worked with GPS data to measure performance on arterials, and the Auckland Transport Authority (2014) updated the measures of reliability and punctuality to incorporate GPS/AVL data in their performance reports.

The tracking and monitoring of bus systems is a topic of great interest all over the world. For example, Cortés, et al. (2011) present the applications of GPS monitoring in diagnosing buses' commercial speed, Berkow, et al. (2007) present the estimation of travel time and other performance metrics using public transportation AVL data, Gokasar & Simseck (2014) use big data for the analysis and improvements of a public transit system, Demiryurek (2016) uses data mining to measure the performance of public transit, and Munizaga (2015) goes further by incorporating the impacts of big data on society in her discussion and studies how the use of big data has the potential to improve not only urban transportation, but also transportation system users' quality of life. Furthermore, other authors have recently developed new technologies to determine performance metrics. For example, Segarra Algueró (2013) discusses the use of smart card technologies to measure the performance of a transits system.

In summary, the analysis of bus systems has been approached by several studies seeking to define measures that calculate performance. Many of them address the problem from an aggregate point of view, tending to measure the quality of service throughout the system. Others discuss the issue by taking a more disaggregated approach, measuring the quality of service at the routes or stops level.

These analyses are done to obtain information to establish the transit system's quality of service, as an elementary part of mobility plans. The TCRP-165 report indicates that a good quality of service can help to retain bus passengers who may have other transportation options (Chaves and Hernández, 2015).

The quality of a system can be established through a performance measure, which refers to any evaluation or comparison measure of a system. According to Hanover, as mentioned in Carter and Lomax (1992), measures that define the performance of a transportation system include effectiveness, efficiency, productivity and service quality. In addition to this, it is also possible to determine the performance of a transportation system by relating the measures of this system, and the use of the service by its users.

Considering only the performance of the transportation system, performance measures can be established that monitor in real-time and can instantly show a bus or route's level of service as a result (Carter and Lomax, 1992). These measurements can be based on data obtained from AVL (Automatic Vehicle Location) and APC (Automatic Passenger Counter) systems.

There are advantages to measuring system conditions in real-time, since travelers can obtain

information about traffic conditions, bus arrival times, and the conditions of all the alternative routes (Berkow *et al.*, 2007). The problem with real-time data lies in the amount of data that needs to be processed to measure performance over time.

Frequency and adherence are among the performance measures recommended by TCRP that can be implemented for real-time monitoring. Frequency measures how often the service is provided. Adherence is used to measure the reliability with which the service adheres to the timeline or commercial headway, plus or minus 10 minutes. In the case of buses, the frequency measures the time in which the units go through a particular stop (TRB, Transportation Research Board, 2012).

Other measures can be estimated through the calculation of scores and service levels. A standardized score can be calculated to determine how many standard deviations an agency's measure is below or above the average (Carter and Lomax, 1992). Levels of service are performance categories that measure the quality of a given service based on expected characteristics and defined by established standards.

Considering the users' use of the service, the performance of a bus system is related to the passengers boarding in the system. In other words, the system's performance, and more specifically, of the buses, can affect the boarding. Therefore, it is important to take into consideration the passengers boarding in a stop and their use of the service, since these are key factors for the justification of the system, given that the system depends on the ridership both at an economic level (passengers pay for a service fee) as on a social level (the service serves everyone equally).

A system with service regularity is attractive to passengers, as their activities will not be affected by unexpected or unjustified delays. In addition to the regularity of service, an increase in vehicle frequency usually increases ridership (Evans et al., 2004). Higher waiting times, associated with an irregular service, may increase the number of passengers left waiting at the stop (captive passengers)and decrease the number of passengers who choose to use the service instead of other travel options or whose itineraries are that cancelled or varied due to service delays.

On the other hand, it must be considered that the boarding of passengers at a stop does not only depend on the transportation system's performance, but also on each passenger's characteristics, the stop, the area surrounding the stop, and the activities developed around it. Gutiérrez et al. (2011) classify these factors into three types: environmental dimensions, socioeconomic factors, and the station's characteristics.

Finally, there are proposed models that try to explain the ridership and the boarding using statistical models, such as the one developed by Chu (2004) which attempts to explain the approach using a Poisson model. Others, like the one developed by Gutiérrez et al. (2011), relate the distance to the stop as a factor associated to boarding.

3 METHODOLOGY

The research methods applied to this project correspond to the category of quantitative methods (Hernández, 2006). The literature related to this work was studied all along the development of this project. Initially, the literature review was conducted to understand the approach that other researchers have taken to examine GPS/AVL big data applications to transit systems performance metrics. Methodologies were also studied to determine how to process the information available. Another vital task was to obtain GPS/AVL data from an actual system. This task was completed by acquiring one month of historical AVL/GPS data for all the in-service buses of the Metropolitan Bus Authority (AMA, for its acronym in Spanish) that serves various municipalities of the Metropolitan Area of San Juan. Initially, the data was cleaned and processed using MATLAB®, but after determining that it was a very complex task to develop from scratch since the methodologies to handle big data scattered in several matrices at the same time, it was decided to switch to Knime® which is a software specifically designed to handle big data. The current performance measures were calculated for one of the AMA routes to ensure that we could process the information available by using this program.

Another track of this project included the stops and their influence areas. The stop, boarding, and alighting data were collected along one of the main routes of the AMA system. After that, Census data related to the influence area of the stops were obtained to characterize the activity system around the stops, and therefore around the corridor, formed by the route under study. A statistical correlation analysis was performed with the data and after that, several models were estimated to select the one with the best fit for the explanatory variables considered. Finally, a new set of performance measures was developed and exemplified with the data available for the transportation system and the activity system in the studied corridor. In the end, conclusions and recommendations were drawn from the whole process.

A summary of the methodology is presented below in Figure 3.

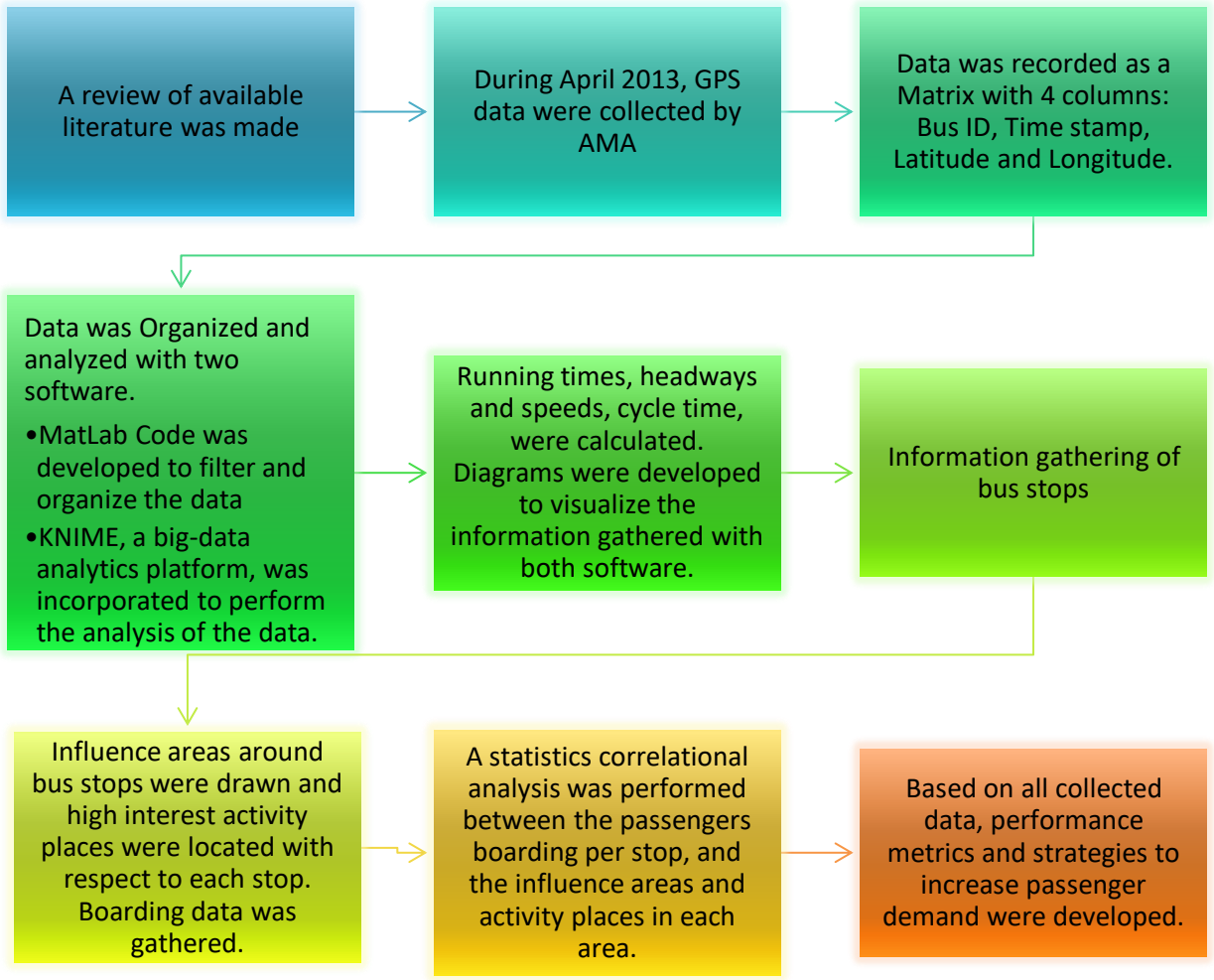


Figure 3 Summary of Methodology

4 DATA

4.1 OPERATIONAL DATA

The operational data was obtained using the GPS/AVL database available from the AMA bus system. This database is comprised of large text files corresponding to the location of all buses and routes. The files are compressed archives of daily operations. The data provided by AMA have a single variable, defined as "timestamp," which contains the following information in a single text string: bus identification, date (in year-month-day format), time (in hour-min-sec format), and the Geographical location in latitude-longitude format. Figure 4 shows a sample of the GPS/AVL data obtained from AMA.

```
817.00    20130401041523.00    18388479.00    -66081631.00
817.00    20130401042103.00    18388460.00    -66081503.00
817.00    20130401042113.00    18388310.00    -66081435.00
817.00    20130401042120.00    18388211.00    -66081590.00
817.00    20130401042126.00    18388178.00    -66081776.00
817.00    20130401042126.00    18388183.00    -66081745.00
817.00    20130401042132.00    18388130.00    -66081970.00
817.00    20130401042137.00    18388071.00    -66082156.00
817.00    20130401042142.00    18387963.00    -66082335.00
817.00    20130401042147.00    18387901.00    -66082516.00
817.00    20130401042211.00    18387863.00    -66082725.00
817.00    20130401042219.00    18387933.00    -66082865.00
817.00    20130401042225.00    18388093.00    -66082938.00
817.00    20130401042228.00    18388205.00    -66082953.00
817.00    20130401042230.00    18388331.00    -66082975.00
817.00    20130401042235.00    18388636.00    -66083036.00
817.00    20130401042240.00    18388921.00    -66083096.00
817.00    20130401042245.00    18389183.00    -66083135.00
817.00    20130401042250.00    18389396.00    -66083088.00
817.00    20130401042255.00    18389545.00    -66082876.00
817.00    20130401042300.00    18389628.00    -66082538.00
817.00    20130401042305.00    18389726.00    -66082086.00
817.00    20130401042310.00    18389755.00    -66081524.00
```

Figure 4 Sample of the Original Data Delivered by AMA

4.2 BUS STOP STUDY

The bus stop study database corresponds to a study conducted by the University of Puerto Rico and partially supported by this project (Cordero, 2015). This work consisted of an evaluation of the stops along Route 5, based on their characteristics and an analysis of boardings and alightings (on-and-off passenger counts). One hundred stops were evaluated, and information was gathered about location, description, and conditions. Figure 5 presents two georeferenced photos. The first one, at the top of the figure, shows all the stops along Route 5. The second one, in the right bottom corner, shows the location of Route 5 and its surrounding areas in the SJMA. The box in the left corner indicates the general statistics for Route T5, which include the following:

(a) length of 24.5 miles;

(b) number of buses assigned: seven (7);

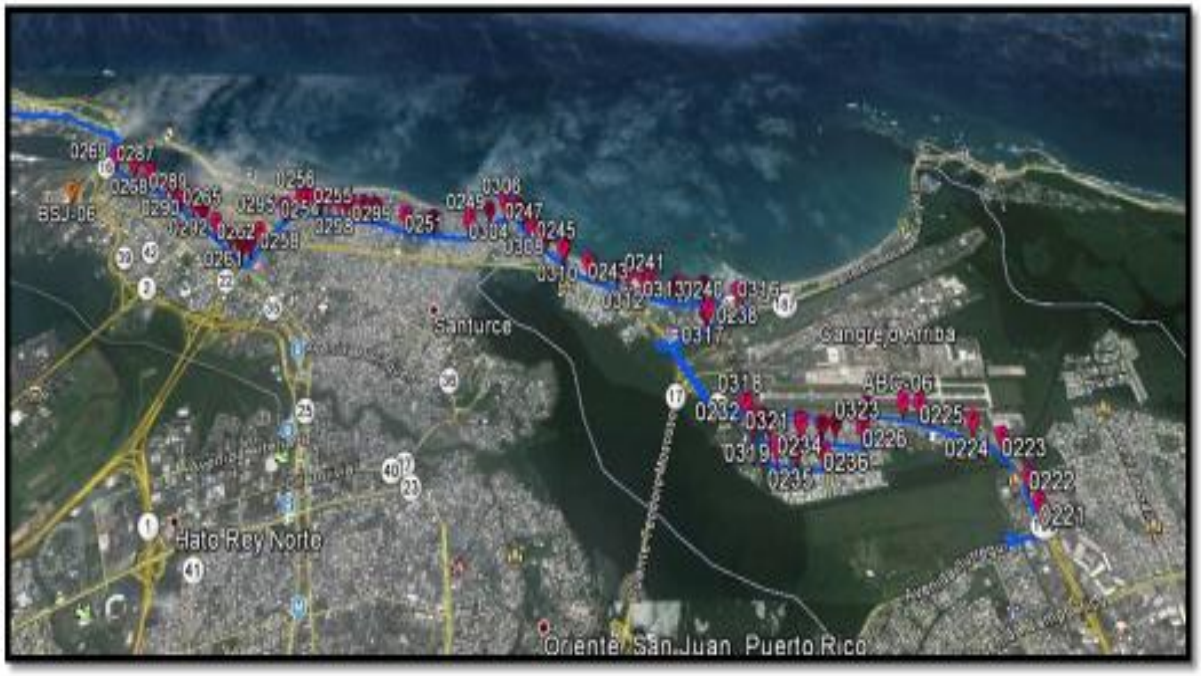
(c) average number of daily passengers on this route: 3,350 pax/day;

(d) municipalities connected by Route T5: San Juan and Carolina; and

(e) a note indicating that the stops located in the Old San Juan islet are not included in this study because at the time of the inventory the area was under construction.

Figure 5 Stops' Location and Description in Route 5 AMA

Figure 6 presents an example of the tile format figures included in the stop study for each one of



the stops along Route T5. The picture on the left of this Figure shows the general condition of the actual stop when the study took place. The box with a yellow background in the top right corner corresponds to the stop number, and the box with a blue background in the middle right corresponds to the location of the stop. Finally, the picture in the bottom right corner shows the georeferenced location of the stop in a photo from Google Maps®. More information related to the stop study is presented in Appendix A.

Figure 6 Example Number and Location of Stops

4.3 ACTIVITY SYSTEM DATA

The activity system database was obtained from various sources including Census data and geographic information that corresponds to the influence area adjacent to each bus stop along the route selected for the analysis. These influence areas are delimited by 0.25 miles around the stops to obtain the demographics of those who live in each area. Figure 7 presents the corridor for AMA Route T5, indicating the influence area of each bus stop along the route.

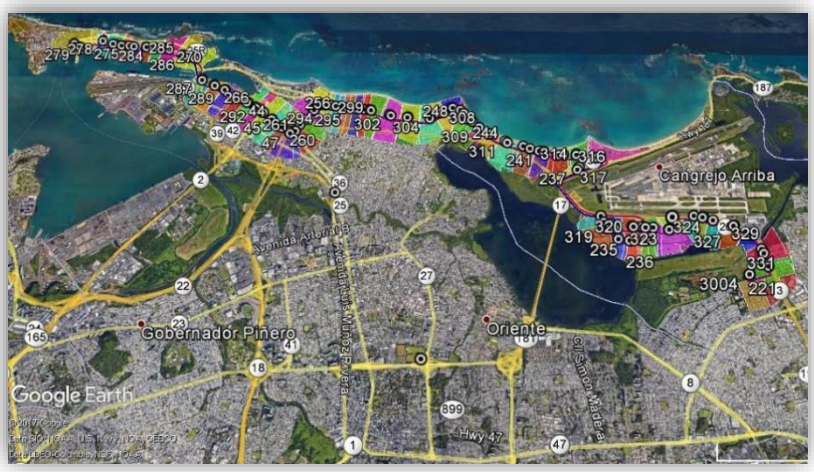


Figure 7 Influence Areas around the Stops on Route 5 AMA

Google Earth was used to define the influence areas around the stops. The rule was to enclose all possible sources or destinations in a polygon within a distance equal to or less than 0.25 miles for each intermediate stop and 0.50 miles for each of two terminals (Covadonga and Iturregui, for Route T5). Then, the scope of the surface for each stop's influence area was calculated. Figure 8 shows more details of a sample of the influence areas for six of the stops along Route T5.

Figure 8 Zoom Showing Details of the Influence Areas around the Stops on Route 5 AMA

As was mentioned, the activity system is represented in this project by trip generators and the sociodemographic data of each stop influence area. Some activity system variables were calculated using the measured area and average rates found in the Census data. Other variables indicated the presence or lack thereof of places that could contribute to add passengers to the boarding/alighting process. For instance, the presence or absence of Hospitals, Schools, Governmental Offices, Industrial Zones, Recreational Zones and Touristic Zones in the area is noted.

The website gis.pr.gov was the source for the following data: population, housing, education, health, tourism, prisons, parks, income, government offices, public squares, industries, recreational places/areas, public housing located around each stop influence areas. This website is the official internet site in Puerto Rico to obtain GIS referenced data. Figure 9 shows a screen capture of this website.

Figure 9 PrintScreen of Website gis.pr.gov

Georeferenced data was essential for visualization purposes. The website gis.pr.gov presents georeferenced data in various topics related to the activity system as shown in Figure 10.

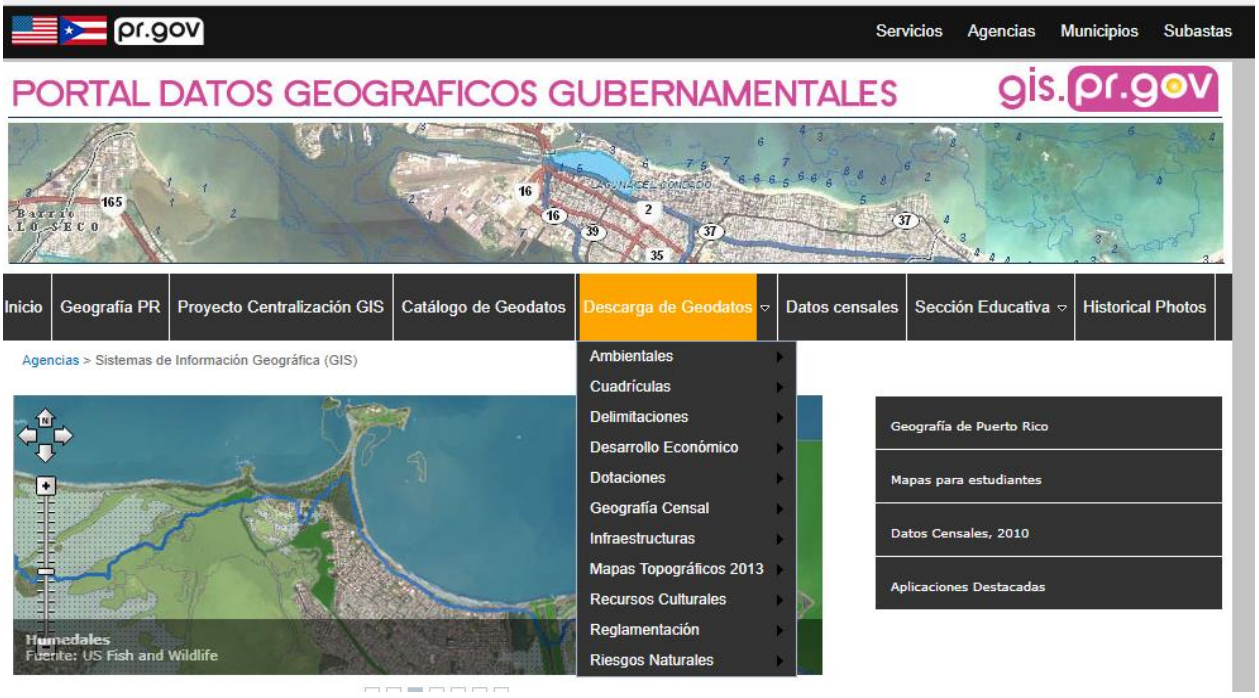
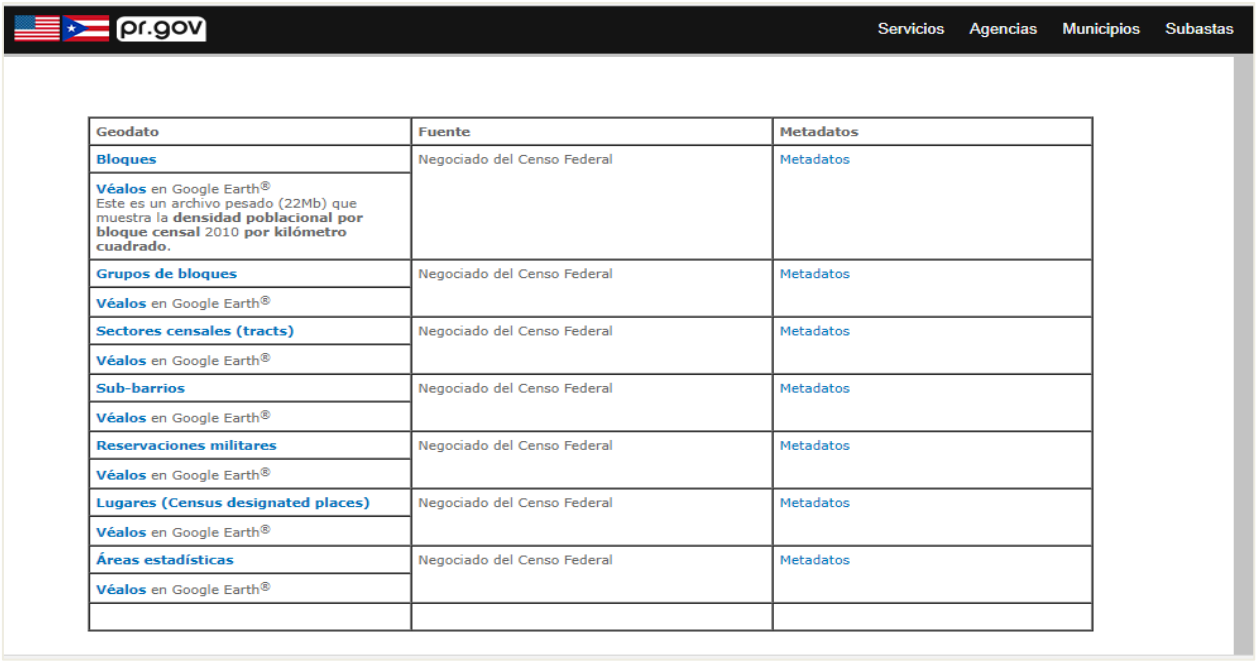


Figure 10 GIS Data of Puerto Rico

Figure 11 shows the details of all the subdivisions that can be used to retrieve Census-based data from the gis.pr.gov website. The data may be displayed in Google Earth® format to make it easy to use. The area unit, called blocks, was selected to retrieve the data because it is the most detailed of all the subdivisions provided in this database. However, in some cases, the delimitation of the blocks did not precisely coincide with the limits of the enclosed influence area. In those cases, additional calculations based on the measured area that belong to each stop's area of influence was needed.



Geodato	Fuente	Metadatos
Bloques	Negociado del Censo Federal	Metadatos
Véalos en Google Earth® Este es un archivo pesado (22Mb) que muestra la **densidad poblacional por bloque censal** 2010 **por kilómetro cuadrado.**		
Grupos de bloques	Negociado del Censo Federal	Metadatos
Véalos en Google Earth®		
Sectores censales (tracts)	Negociado del Censo Federal	Metadatos
Véalos en Google Earth®		
Sub-barrios	Negociado del Censo Federal	Metadatos
Véalos en Google Earth®		
Reservaciones militares	Negociado del Censo Federal	Metadatos
Véalos en Google Earth®		
Lugares (Census designated places)	Negociado del Censo Federal	Metadatos
Véalos en Google Earth®		
Áreas estadísticas	Negociado del Censo Federal	Metadatos
Véalos en Google Earth®		

Figure 11 Available Gis Data

Source:http://www2.pr.gov/agencias/gis/descargaGeodatos/GeografiaCensal/Pages/GEODATOS-CENSO-2010.aspx

In addition to data related to the population, other data were also taken from the gis.pr.gov website. The additional data included the number of places adding passengers to the bus stop as well as the distance between these places and the stops under analysis. Also, the data from the stops study was used to determine the approach and overflow of passengers at each stop. All these data were combined to obtain a better representation of the activity system around the Route T5 bus corridor and its influence area.

5 CURRENT PERFORMANCE MEASURES

The estimation of current performance measures was the first step before continue developing new metrics. Two procedures are presented. The first one is based on a Matlab® where series of subroutines were designed to organize the data in matrix format to take advantage of Matlab® tools. However, once the data was clean and prepared by bus-route, the whole method of managing big data using Matlab® was too elaborated and time-consuming. The development of new performance measures was delayed by the long hours required to process the information. Therefore, a big-data management program called Knime® was used to continue with the new developments. Both processes are presented in this section.

The Knime® software proved to be more versatile in managing the data. The previous experience with Matlab® was useful to generate new algorithms to clean the data and obtain data sets separated by bus-routes. After that, the current performance metrics were developed as presented in section 5.1.2.

5.1 ESTIMATION OF CURRENT PERFORMANCE MEASURES USING MATLAB

Matlab® was used initially to estimate current performance metrics using big-data. The raw data obtained from the AMA's AVL system required cleaning and pre-processing before a set of data separated by bus-routes was available to calculate the current performance metrics. This section of the report presents the pre-processing, and the application of the algorithms developed to clean the data. After that, a series of figures demonstrate the procedures used to calculate the current performance metrics.

5.1.1 METHODOLOGY

The proposed methodology has three stages. The first stage was to gather the GPS/AVL files from the AMA Control Center to obtain the information on the location of the vehicles in real-time. The software originally used by AMA was able to display the location of the buses on real-time onto a set of big screens located in the Control Center at the AMA headquarters. The information was also displayed in local computer monitors using software accessible over the Internet. The information about stops along each route was digitalized and saved onto a virtual database from which the software operated. It is relevant to indicate that, at the time that the data was provided, AMA had a private company under contract to manage the information processing aspects of their GPS/AVL system. They were able to retrieve the data that we are using in this study, but AMA discontinued the data collection service after a couple of months. Recently AMA is working to continue their vehicle tracking program.

The second stage was the analysis of AMA performance with at least a month of archived data. The purpose of this stage was to translate raw data into manageable figures. The analysis helped establish which specific metrics AMA would need to record continuously to assess its performance.

The data of AMA buses was organized by day in text format. For presentation in this report, only small samples of the data are included in the Figures shown. The data could not be imported directly to MATLAB® because of the format that was used to write the information in the data files. As shown in Figure 12, the data were separated only by a comma, which makes it difficult to work in MATLAB® and to recognize these data as an array.

```
1025,20130401001217,18387983,-66081807
1024,20130401020251,18388298,-66082153
928,20130401024958,18387785,-66081091
843,20130401025121,18388551,-66082674
```

Figure 12 AMA Data by Day in Text Format

To start with the iterations, a small sample was processed using Microsoft Office Excel® to change the format (See Figure 13). Afterwards, a subroutine was developed to read the original file and write the working file in a format suitable for working on MATLAB®. Having solved this problem, the data was organized in matrix format to be used in MATLAB®.

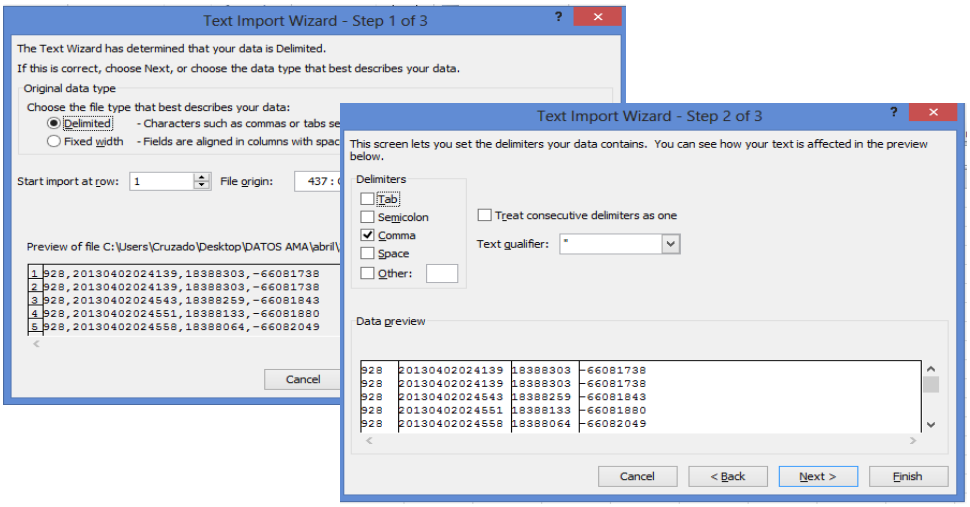


Figure 13 Use of Excel Program

The next obstacle encountered was how the coordinates were recorded. While they should have been in the order of tens, they were in the order of millions. With the coordinates in the correct format, the next step was to determine the route associated with each bus identified in the data file. A series of algorithms were developed considering the GPS coordinates of the primary transfer facilities in the transit network of the SJMA (See Figure 14, Figure 15 and Figure 16). A rectangle

considering the coordinates around each transfer facility and each stop along the bus routes was used as a reference to determine the route of each bus in the database. Once the route was identified, the data was plotted to build the space-time diagrams corresponding to each bus per route. The current AMA performance measures were calculated using a series of algorithms developed for each metric.

Figure 14 Station Rio Piedras

Figure 15 Station Iturregui

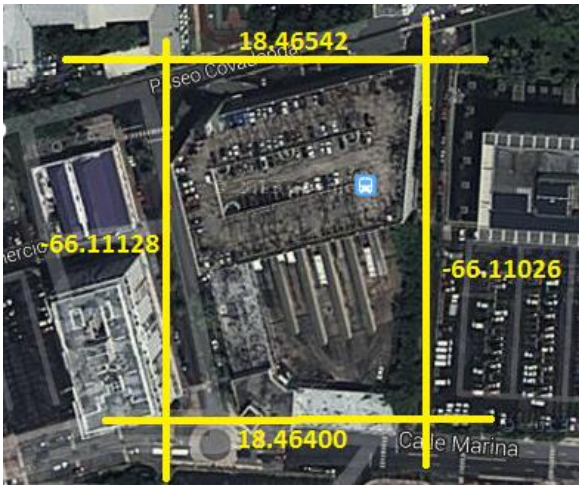


Figure 16 Covadonga Station

5.1.2 PROBLEM WITH THE TECHNOLOGICAL TOOLS FOR COLLECTING DATA FROM AUTOMATIC VEHICLE LOCATION (AVL, GPS).

The original aim of our project was to develop performance measures that could be implemented in real-time. At the time we started the project, AMA still had an AVL/GPS system working through a private company that used to handle both the equipment instalation and maintenance in their vehicles and the data collection and display. They provided the data that we used to develop the research project presented in this report. However, due to financial constraints in the agency, the AVL system was disabled in 2014 and has not been reestablished yet. An attempt was made to develop an in-house system that was installed in some of the vehicles. The system developed was initially successful, but very fast run also in trouble because of similar financial constraints that were encountered before. Currently, the Agency is working to put in place a new system to handle the automation of all the planning, operations and administration of the service.

As a result of the lack of a working real-time AVL system, this report presents all the developments using the one-month historical data that was obtained from AMA at the beginning of our project.

5.1.3 PREPARATION OF ROUTES AND STOPS DATA

During the period of performance of this study, the AMA service network experienced various changes. Some of the service routes suffered changes to streamline the operation, others were consolidated, and yet others eliminated. Therefore, some changes were made to our project, to consider the data of those routes that remained unchanged or suffer minor changes only. The final routes selected for further study were digitized using Google Earth® for display purposes. Figure 17 presents the path and stop information for AMA route 2 that suffer only minor changes in stop location.

Figure 17 Location of the Stops

A database per route was created to improve the program functionality. The database included a list of the stops in sequential order, their coordinates, and the accumulated distance from one of its terminals.

5.1.4 PREPARATION OF PERFORMANCE MEASURES

A list of possible performance measures was prepared to make operational recommendations to AMA. The list includes the following performance measures:

- Frequency or Interval
- Percent Person-Minutes Served
- Transit Service Accessibility Index
- Transit Accessibility Index
- Local Index of Transit Availability
- Index of Transit Service Availability
- On-Time Performance (Fixed-Route)
- Headway Regularity or Adherence
- Run-Time Ratio
- Travel Time

• Travel Time Variability

• Transit-Auto Travel Time Ratio

• Reliability Factor

• Route Directness

• Delay

• Travel Speed

• Percent of Buses Exceeding the Speed Limit

It should be noted that most of these measures are suitable for long-term planning or monthly or annual analysis. The only ones that could be used in real-time are the following:

• On-Time Performance (Fixed-Route)

• Headway Regularity or Adherence

• Run-Time Ratio and

• Travel Time

5.1.5 RESULTS OF THE INITIAL ANALYSIS USING MATLAB

As previously indicated, one month of AVL data from main routes in service was available for the initial analysis. The results show the analysis of the running time of the buses, the headway variation, and adherence, from which it is possible to calculate many of the underlying performance metrics.

Figure 18 shows an example of the Space-Time diagram developed from the data of Route 2 (one of the main AMA routes), which has the ability of visually representing: running speed, running times, bunching frequency and headway variation, among other classical performance measures.

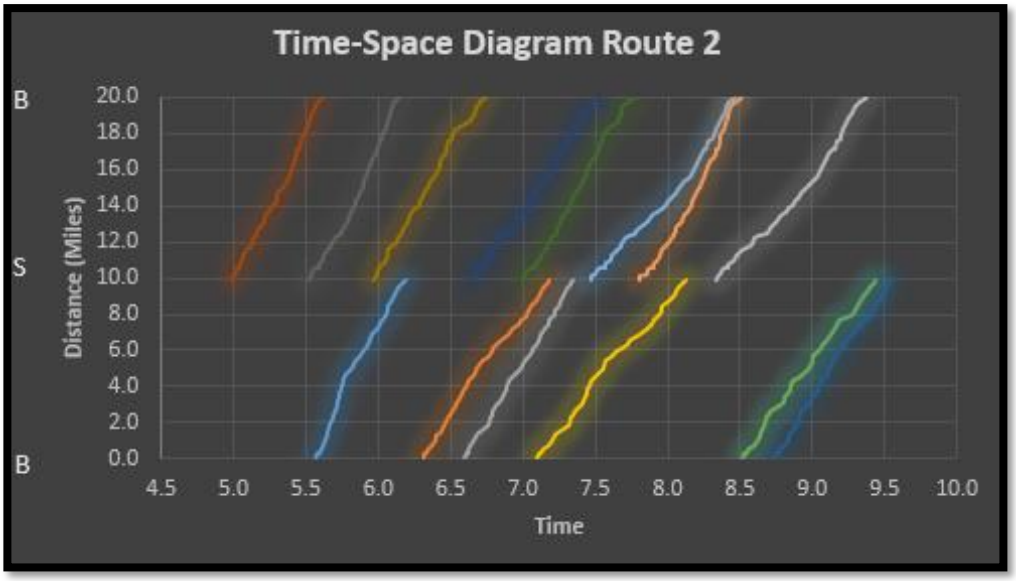


Figure 18 Example of the Space-Time Diagram for Route 2

Figure 19 shows hourly variations of the running times during the day, and Figure 20 shows the variety of on-route time throughout the day. Finally, Table 1 presents an example of the descriptive statistics of the running time and headways.

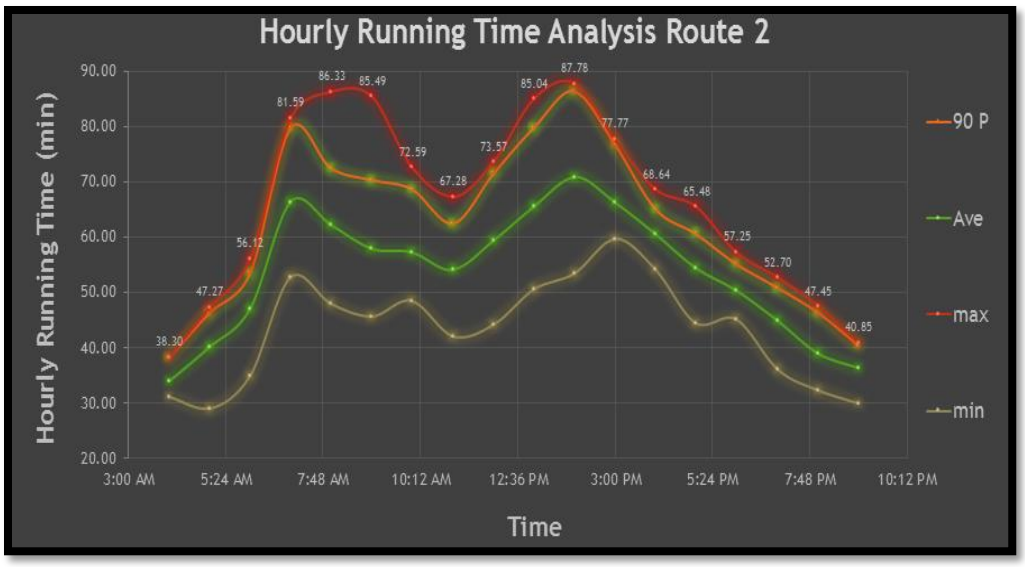


Figure 19 The Variation of Running Times, Hourly Mean

Figure 20 Variation of Route 2 throughout the Day

Table 1 Descriptive Statistics of the Running Time and Headways

Route Segment	Running Time				Headway			
	Mean	85th Percentile	Max	Min	Mean	85th Percentile	Max	Min
Route 2 Bayamón-Santurce	51	64	87	29	36	51	86	1
Route 2 Santurce- Bayamón	57	72	156	32	34	52	102	1
Route 3 Río Piedras-Cataño	53	67	79	26	38	55	124	10
Route 3 Cataño-Río Piedras	53	63	72	30	37	51	100	13
Route 5 Iturregui-San Juan	58	67	95	35	43	65	129	8
Route 5 San Juan-Iturregui	51	62	84	33	42	65	134	7
Route 6 Iturregui-Carolina	27	33	43	12	53	66	394	14
Route 6 Carolina-Iturregui	29	36	72	13	52	70	377	7
Route 7 (Both ways)	72	88	118	48	67	108	138	20

5.2 ESTIMATION OF CURRENT PERFORMANCE MEASURES, USING KNIME SOFTWARE

The methodology used for processing and analysis of the data presented in this section was based on the large data mining model CRISP-DM (Cross-Industry Standard Process for Data mining) incorporated in the software named Knime®. The CRISP-DM methodology, as shown in Figure 21 is divided into stages that show the life cycle of data mining. These stages are:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Assessment
6. Implementation

The CRISP-DM methodology is flexible and can be easily customized for the project where it is used, allowing to create a data mining model that fits particular needs (Corporation, 2011). This section presents the stages of the CRISP-DM methodology applied to the analysis of the AMA database.

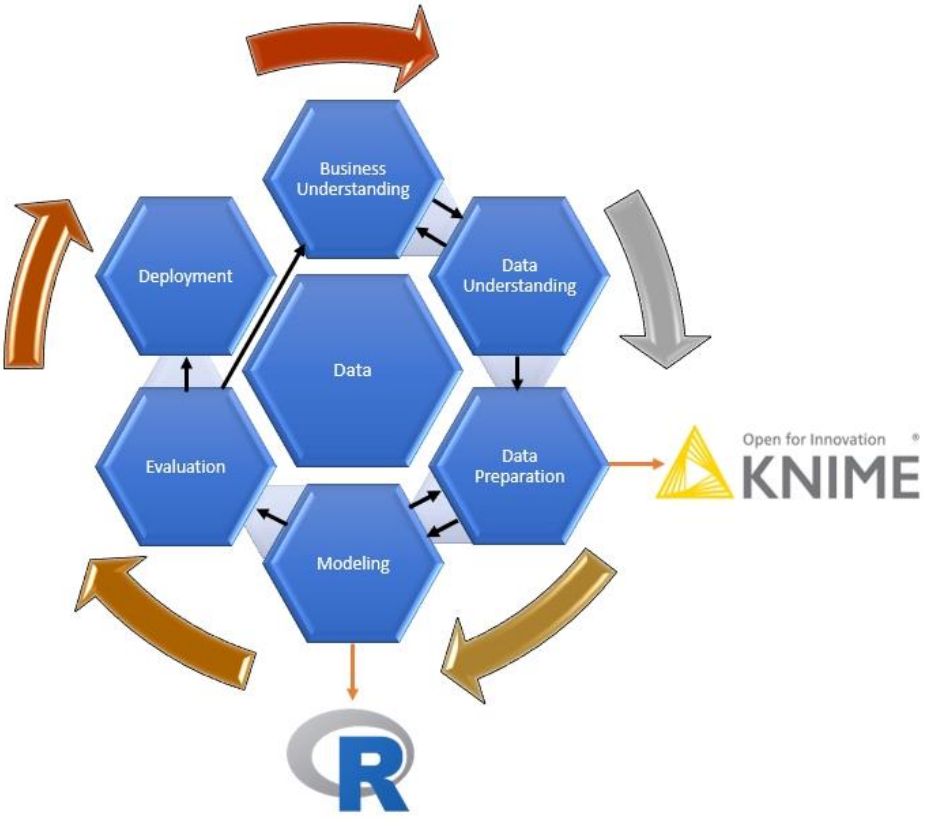


Figure 21 Cross-Industry Standard Process for Data Mining

5.2.1 BUSINESS UNDERSTANDING

The analogy in CRISP-DM terms of the "business" being studied here corresponds to the Metropolitan Bus Authority (AMA). It is the largest bus system in Puerto Rico in terms of routes, budget, fleet, and sponsorship. The system comprises 30 fixed routes and 5 incentive program routes, for disadvantaged populations, distributed as follows:

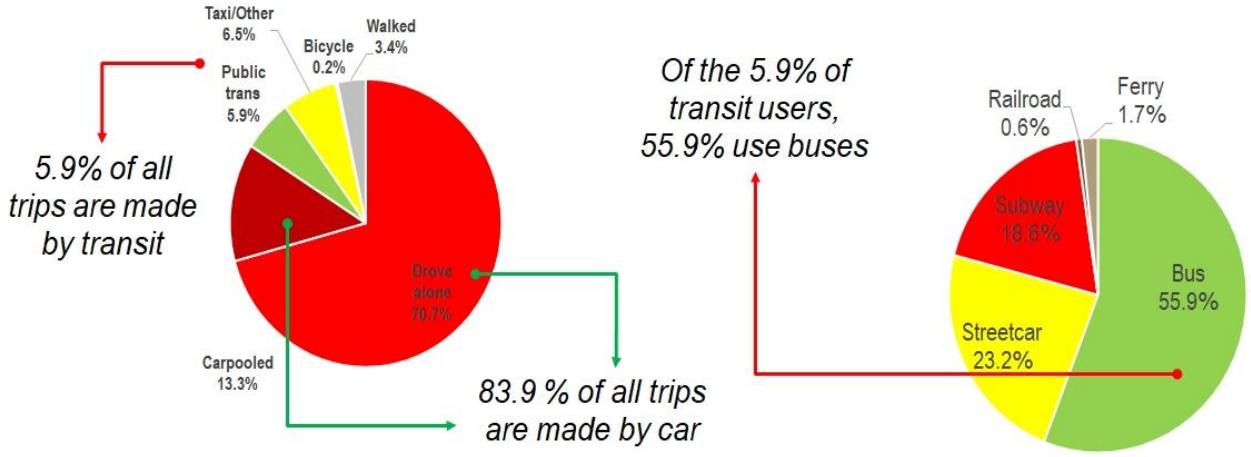
• 3 Express routes with frequencies between 10 and 30 min

• 10 Main routes with frequencies between 10 and 25 min

• 7 Circulation routes with frequencies between 20 and 30 min

• 10 Distribution routes with frequencies between 30 and 90 min

In addition to the fixed routes, the system has 10 terminals and the service "Llame y Viaje" (LV), a paratransit bus system dial-a-ride inaugurated in 1992 for customers with disabilities or special needs. It complements AMA by providing a more direct and far-reaching service aimed at passengers with disabilities and the elderly. The AMA system serves a population of approximately 1,120,859 inhabitants who work and live in the core municipalities of the San Juan Metropolitan Area. Of all the trips made in the metropolitan area, only 5.9% are made by transit, and from this percentage, only 55.9% use buses (Pipicano *et al.*, 2016). The Metropolitan Bus Authority (AMA) manages most of this 55.9%.



(a) Mode choice *(b) Transit mode choice*

Figure 22 San Juan Mode Choice

Source: TIGER Database from TransCad. (2)

According to Navarro Díaz (2017) , whose work coincides with that Pipicano et al. (2016), only 2.7% of San Juan's urban population uses public transportation (note that it is approximately 55.9% of 5.9%) (as quoted in Rodríguez Marrero, 2016). In addition, 3.5% walk and less than 1% ride bicycles. The municipalities with the highest use of public transportation are San Juan and Cataño, where 9.1% of the population use this service.

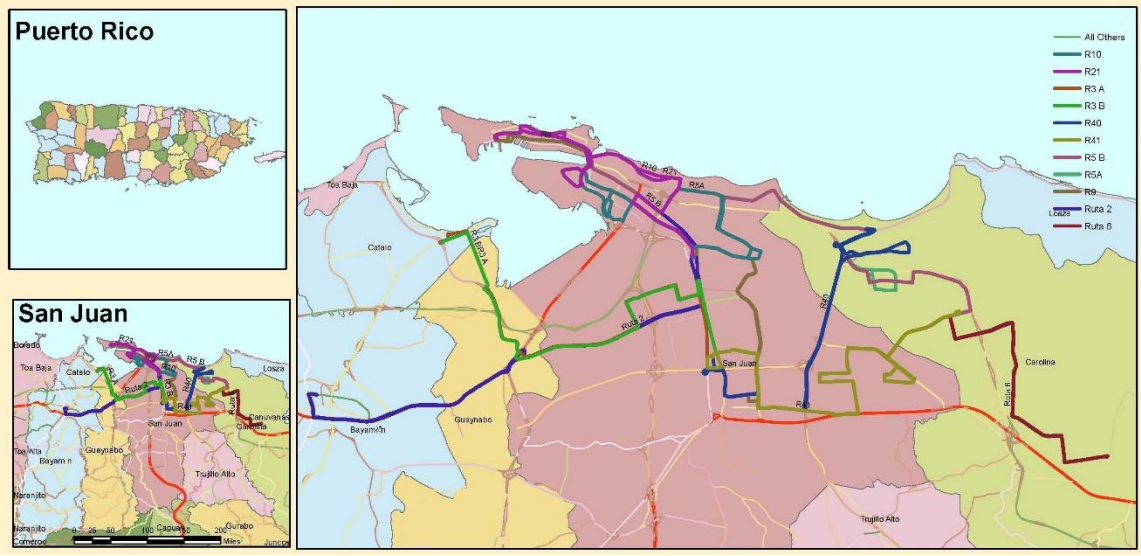


Figure 23, shows a distribution of the 30 routes managed by AMA. These routes, in addition to the MetroBus I and II and the Express 91, 92, are presented below.

Figure 23 Routes of Metropolitan Bus Authority

The methodologies developed in this project are presented in detail for Route 5 of the AMA system (See Figure 24). This is a regular route of the system that provides service and connect very important activity locations in the San Juan Metropolitan Area. Route 5 provides service between San Juan and Carolina. It runs form the Covadonga Terminal (3006) in San Juan to the Iturregui Terminal (3004) in Carolina. It also provides service to Santurce and Isla Verde along 25 miles in both directions with 117 stops. According to the 2015 data, Route 5 has an average ridership of 3,300 passengers per day (Cordero, 2015). Figure 24 shows the end terminals and all the intermediate stops along Route 5.

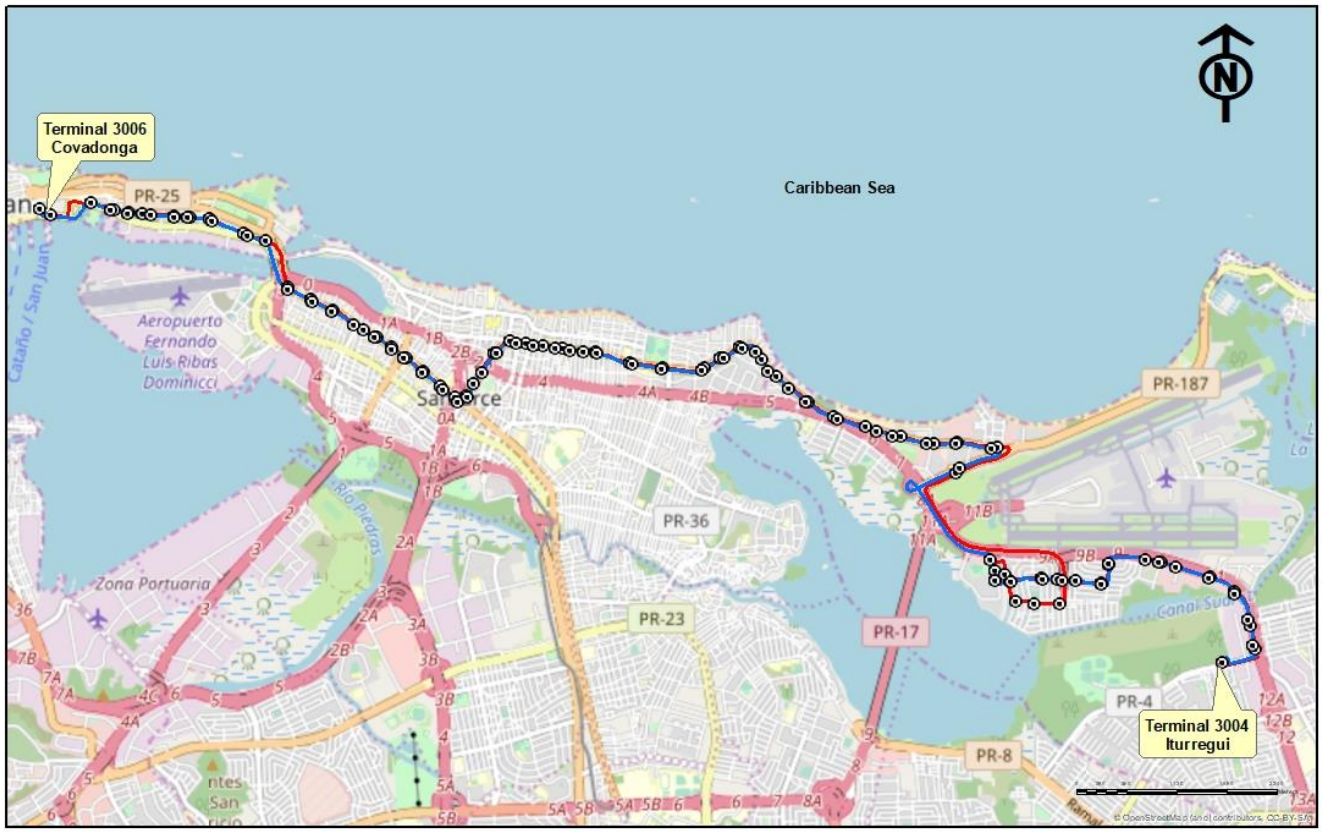


Figure 24 Trajectory of Route 5 with Terminals and Stops

5.2.2 DATA UNDERSTANDING

This stage of the CRISP-DM methodology is presented in this chapter to have a complete discussion of all the stages. However, this section presents only a summary of what has been already presented in previous sections related to the data used for analysis purposes.

The data were obtained from the following three different databases: (1) AMA database, (2) Bus Stop database, and (3) Census database. (1) **The AMA database** corresponds to large text files comprising the location of all buses and routes, as described in section 4.1. (2) The **Bus Stop database** corresponds to the AMA bus stops inventory and evaluation conducted by the University of Puerto Rico, as described in section 4.2. The variables obtained from this study that will be included in the analysis are the boarding of passengers in each bus stop in a 6.5-hour period during a typical working day in March 2015 from 6:30AM to 1:00PM, as mentioned in section 7.1. (3) **The Census database** corresponds to the Census data in the area of influence of the bus stops along the route selected for analysis, as described in section 4.3.

The final database that was used for analysis and modeling was a combination of these three databases considering the influence area of the bus stops along the AMA Route 5 shown in Figure 25.

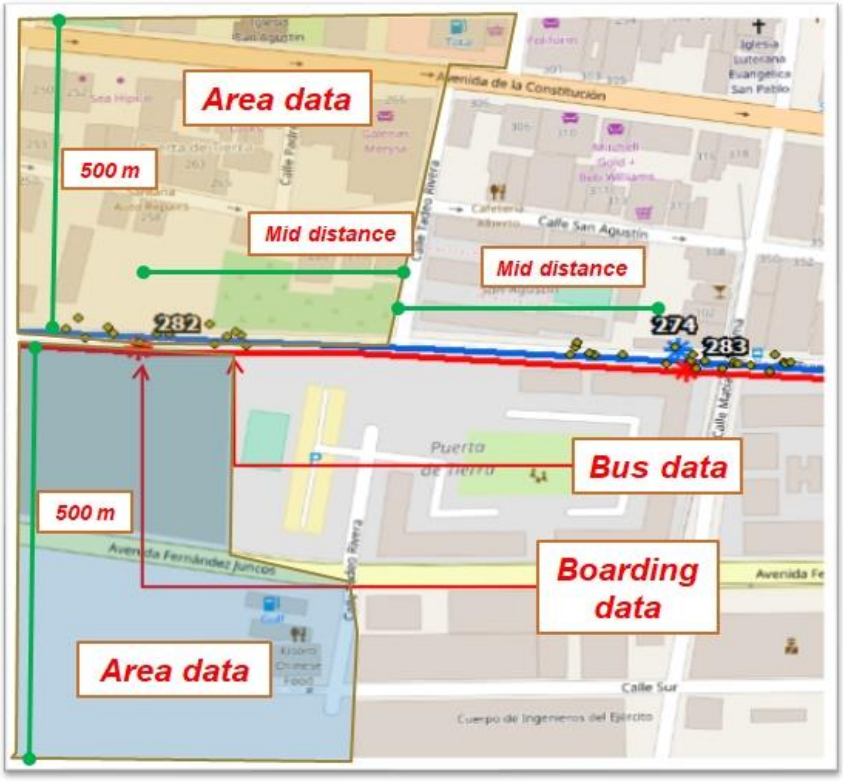


Figure 25 Bus Stop Area

The variables selected for the analysis are: Average Income per capita measured in thousands of dollars per year, and Mid distance between two consecutive stops measured in meters. The variables hospitals in the area, schools in the area, governmental offices in the area, industries in the area, recreation in the area and tourism zones in the area, are dichotomous variables measured as 1 or 0 representing the presence (1) o absence (0) of this type of travel generators in the zone. These variables were grouped as shown in Table 2.

Table 2 Description of Variables

Variable	Response
Bus	
Performance Measure (Headway, Score, etc)	Numeric - Continuous
Passengers	
Boarding	Numeric - Discrete
Bus Stop Influence Area	
Average Income per capita	Numeric - Discrete
Mid distance between stops	Numeric - Discrete
Hospitals in the area	Numeric - Binary
Schools in the area	Numeric - Binary
Governmental offices in the area	Numeric - Binary
Industries in the area	Numeric - Binary
Recreation in the area	Numeric - Binary
Tourism zones in the area	Numeric - Binary

5.2.3 DATA EXPLORATION AND PREPARATION, DATA PROCESSING AND DATA MODELING

This part of the process was done using two informatic packages for data analysis: Knime ANALITYCS® and R®.

Knime ® is an open source program that uses modular nodes, linked and integrated to effectively improve the data mining process, as shown in Figure 26. The most important aspect of this software is its graphical interfaces that can represent networks of linear or cyclical processes that include the processing, modeling, analysis, and visualization of data. In the case of this project, Knime® was used for the mining process, which includes the cleaning, organizing, and processing of GPS data for all buses, all routes, and all the days selected for analysis.

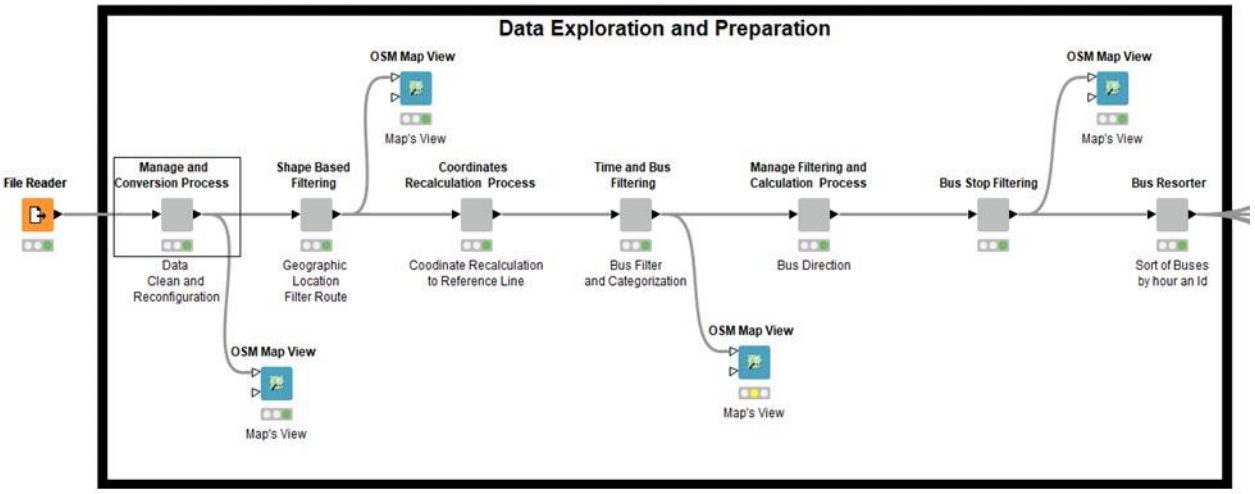


Figure 26 Data Mining Process with Knime Analytics: Data Exploration and Preparation

R® is a statistical computing software package. According to (Faraday, 2009), R® has 4 characteristics that make it one of the best software packages for statistical analysis: Versatility, Interactivity, Freedom and Popularity.

• Versatility: R® apart from being a program, is also a programming language, which is not limited by pre-programmed package procedures.

• Interactivity: Data analysis is inherently interactive, allowing changes based on what we see during the analysis.

• Freedom: R® is an open source software and can be obtained for free. It is compatible with other data management software, such as Knime® and Latex®.

• Popularity: R® is a very popular software for researchers using statistics. R® was used in this work, in conjunction with Knime®, for statistical data modeling.

5.2.4 DATA EXPLORATION AND PREPARATION

The first part of this sub-process corresponds to the exploration and preparation of the data, to make sure that it is adequate for processing and analysis. The data obtained after the initial process corresponds to the separation of the variable "time stamp" in 5 variables named: Date, Hour, nBus, xLon and yLat. Table 3 shows a sample of the K-nime® table resulting from the preparation process based on the original data. The variables shown in Table 3, correspond to the date of the file in which the data were collected, the hour of the GPS' reading signal, the ID of the bus emitting the data, and the geographic coordinate longitude and latitude respectively.

Table 3 Result of the Initial Process of the Data from "time stamp"

Date	Time	nBus	yLat	xLon
01.Apr.2013	06:00:04	817	18.422	-65.992
01.Apr.2013	06:00:20	817	18.422	-65.992
01.Apr.2013	06:00:25	817	18.422	-65.992
01.Apr.2013	06:00:29	817	18.422	-65.992
01.Apr.2013	06:00:30	817	18.422	-65.993
01.Apr.2013	06:00:35	817	18.422	-65.993
01.Apr.2013	06:00:40	817	18.421	-65.993
01.Apr.2013	06:00:49	817	18.421	-65.992
01.Apr.2013	06:00:54	817	18.421	-65.992
01.Apr.2013	06:00:59	817	18.421	-65.992

Using coordinate location, the preprocessed data can be filtered by bus stops along Route 5, as shown in Figure 26. Table 4 presents a sample of the results of this filtering process. The information presents Time, Bus Number and Stop Location along Route 5. Current operation performance measures for route and system performance can be calculated with the time and location of each bus and each bus stop in the route.

Table 4 Buses Filtered by Bus Stop, Based on Its Location and the Bus Stop Location

Time	yLat	xLon	nBus	BusWay	nStop	wStop
06:37:25	18.431	-66.015	817	1	232	1
06:37:30	18.43	-66.015	817	1	232	1
06:37:37	18.43	-66.015	817	1	232	1
06:37:42	18.43	-66.015	817	1	232	1
06:38:55	18.429	-66.014	817	1	231	1
06:39:00	18.429	-66.013	817	1	231	1
06:39:25	18.429	-66.01	817	1	230	1
06:39:41	18.429	-66.01	817	1	230	1
06:39:46	18.429	-66.01	817	1	230	1
06:39:56	18.429	-66.009	817	1	229	1
06:40:31	18.429	-66.007	817	1	228	1
06:40:35	18.429	-66.007	817	1	228	1
06:40:40	18.429	-66.007	817	1	228	1
06:40:45	18.429	-66.006	817	1	228	1
06:41:00	18.429	-66.004	817	1	227	1
06:41:05	18.429	-66.004	817	1	227	1
06:41:22	18.429	-66.004	817	1	227	1
06:41:27	18.429	-66.004	817	1	227	1
06:41:53	18.431	-66.004	817	1	226	1
06:41:58	18.431	-66.004	817	1	226	1
06:42:03	18.431	-66.003	817	1	226	1
06:43:09	18.431	-65.999	817	1	225	1
06:43:39	18.43	-65.994	817	1	224	1
06:43:59	18.428	-65.991	817	1	223	1
06:45:00	18.425	-65.989	817	1	222	1

5.2.5 DATA PROCESSING

The second part of this sub-process, that follows the Data Exploration and Preparation shown in Figure 26, includes the estimation of the usual operational performance measures. The calculated performance measures, as mentioned in section 5.1.4, are the Headway, Average Speed, Cycle Time, Regularity Index, and Coefficient of The variables selected for the analysis are: Average Income per capita measured in thousands of dollars per year, and Mid distance between two consecutive stops measured in meters. The variables hospitals in the area, schools in the area, governmental offices in the area, industries in the area, recreation in the area and tourism zones in the area, are dichotomous variables measured as 1 or 0 representing the presence (1) o absence (0) of this type of travel generators in the zone. These variables were grouped as shown in Table 2.

Table 2. In addition, Space-Time diagram, Speed Histogram, Speed Box-Plot, and Cycle Scatter Plot were drawn.

A new set of performance measures were proposed with a methodology to estimate the performance of a route based on the level of service concept. One of these measures was correlated with all other variables involved and the results are shown in section 7.3.

The third part of this sub-process includes a descriptive analysis of the data, a correlation analysis and the modeling of the data for estimation purposes. This part started with the integration of the three databases obtained in a matrix of 117 records (corresponding to the number of stops) by 8 variables, shown in Table 2. Figure 27 show the statistical methodology used to complete the analysis of this third sub-process. Once the database was consolidated, a descriptive analysis of the data and a correlation analysis of the variables were made. The descriptive analysis was done in order to define the behavior of the data, verify the presence of inconsistencies that could arise due to mining and its possible solutions. The correlation analysis was carried out to identify possible relationships between the explanatory variables that could generate multicollinearity problems. The model estimation was performed to determine how the explanatory variables are related to the response variable. In this case, the response variable is Boardings. The response and explanatory variables considered as part of this sub-process are presented in The variables selected for the analysis are: Average Income per capita measured in thousands of dollars per year, and Mid distance between two consecutive stops measured in meters. The variables hospitals in the area, schools in the area, governmental offices in the area, industries in the area, recreation in the area and tourism zones in the area, are dichotomous variables measured as 1 or 0 representing the presence (1) o absence (0) of this type of travel generators in the zone. These variables were grouped as shown in Table 2.

Table 2.

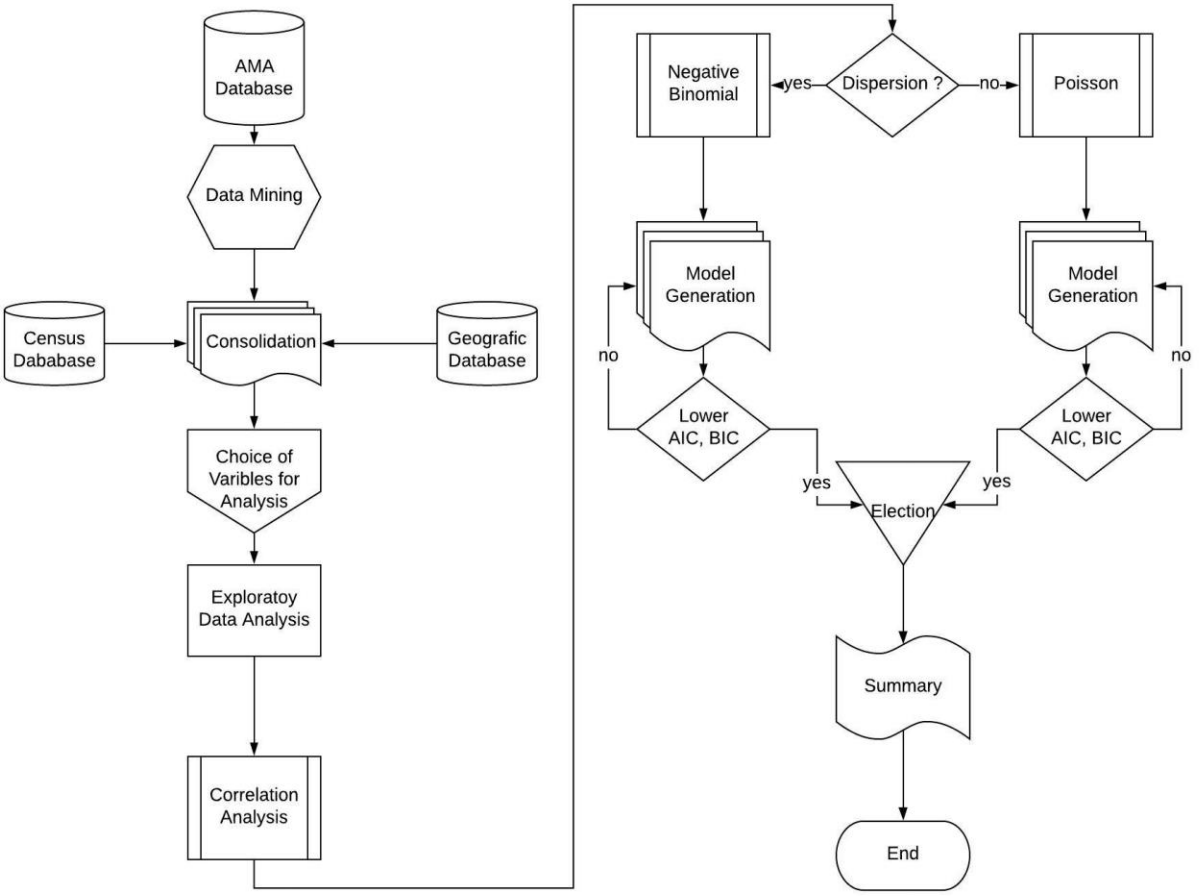


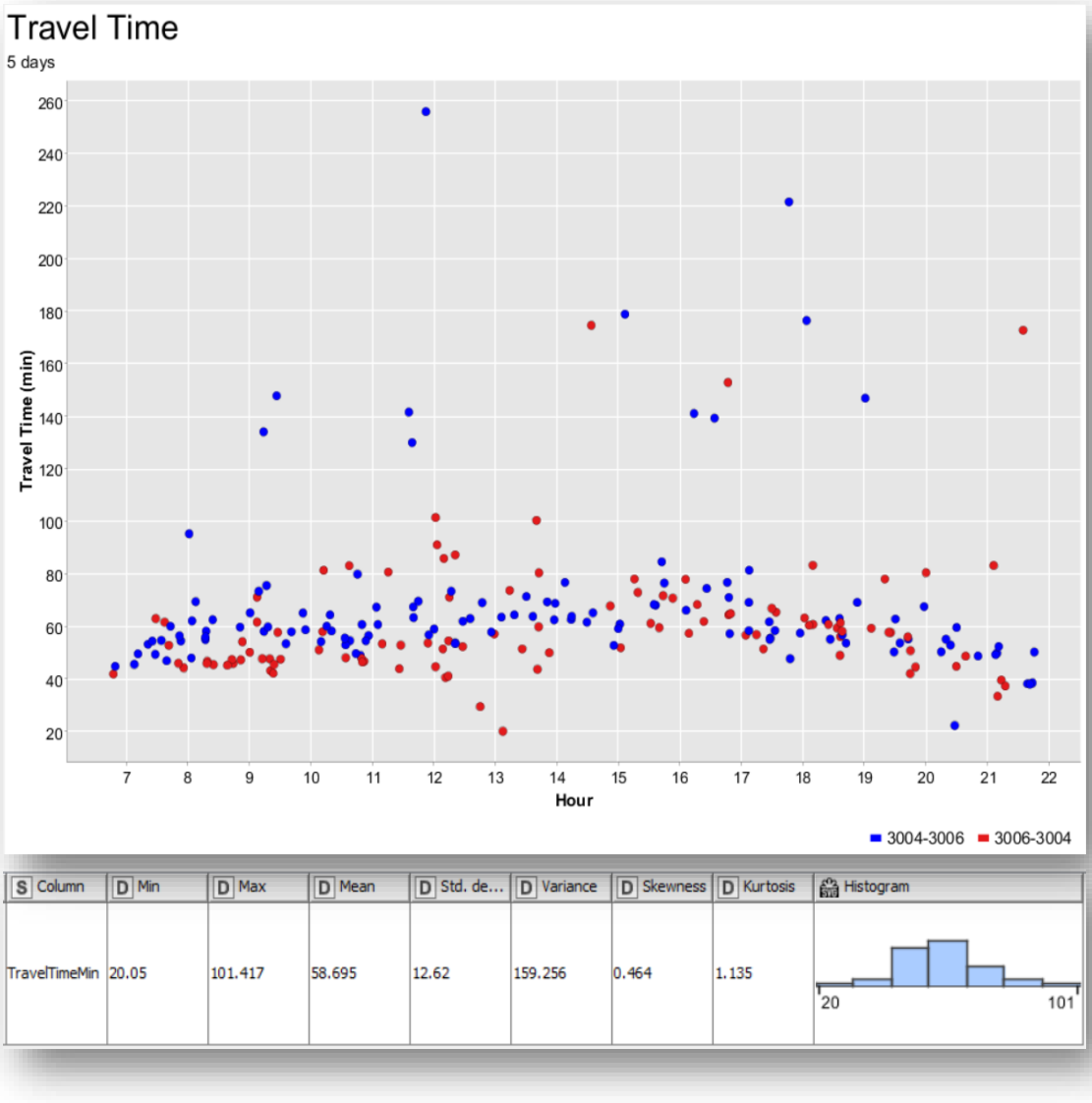
Figure 27 Statistical Methodology

5.2.6 USUAL OPERATIONAL PERFORMANCE MEASURES

Transit systems typically use performance measures to assess their operations. This section presents a group of perfromance measures calculated by processing the big data corresponding to AMA Route 5. The performance measures presented include Travel Time, Running Time, Terminal Time, Cycle Time, Travel Speed, Running Speed, Headway Regularity Index and Coefficient of Mean Variation.

5.2.6.1 TRAVEL TIME

The travel time is defined as “the time necessary to traverse a route between any two points of interest” ((FHWA), 2018). Figure 28 show a scatter plot of a five-day sample of the travel time of each bus and the travel time general descriptive statistics respectively from terminal 3004 (Iturregui) to terminal 3006 (Covadonga) and backward. The statistics were calculated after eliminating the outliers.



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
TravelTimeMin	20.05	101.417	58.695	12.62	159.256	0.464	1.135	20 101

Figure 28 Two-way travel time 3004 to 3006 and backward

Figure 28 shows that the average travel time is approximately 58.7 minutes, in a range of 20.0 minutes to 101.4 minutes.

The scatter plots show several points located far away from the trend. The points in the lower end are unfeasible because they would represent bus speeds that are not possible in the system. Those

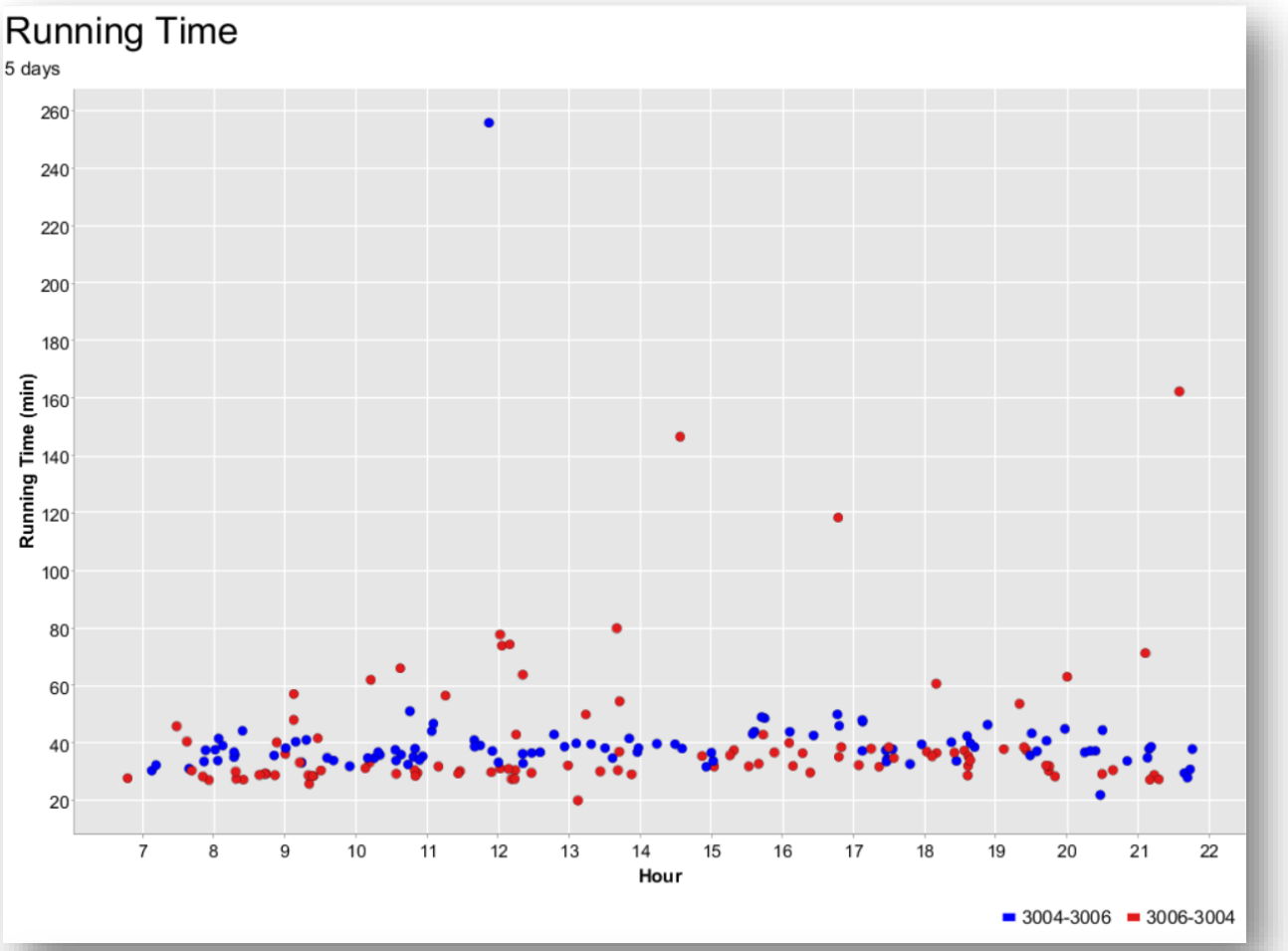
may be errors in transmission or reception of the GPS equipment data or very short runs that are not part of the route, even though the bus was out of the terminal for a short period. Points in the upper end may reflect buses that have mechanical difficulties along the route and may have required mechanical services or even towing to return to a terminal or the main yard for maintenance. Therefore, those values are not used for statistical analysis.

5.2.6.2 RUNNING TIME

Running time refers to the number of scheduled minutes assigned to a bus for moving from one time point location to the next. Running times are accurate when they are sensitive to the varying traffic conditions and passenger volumes over the course of a service day (Transportation Research Board, 1998).

Error! Reference source not found. presents a scatter plot of a five-day sample of the running time of each bus and the general descriptive statistics respectively from terminal 3004 (Iturregui) to terminal 3006 (Covadonga) and backward. An outliers' analysis was performed and the extreme values of running times were not considered to calculate the descriptive statistics.

The average running time was 37.64 minutes in a range of 19.9 minutes to 79.9 minutes.



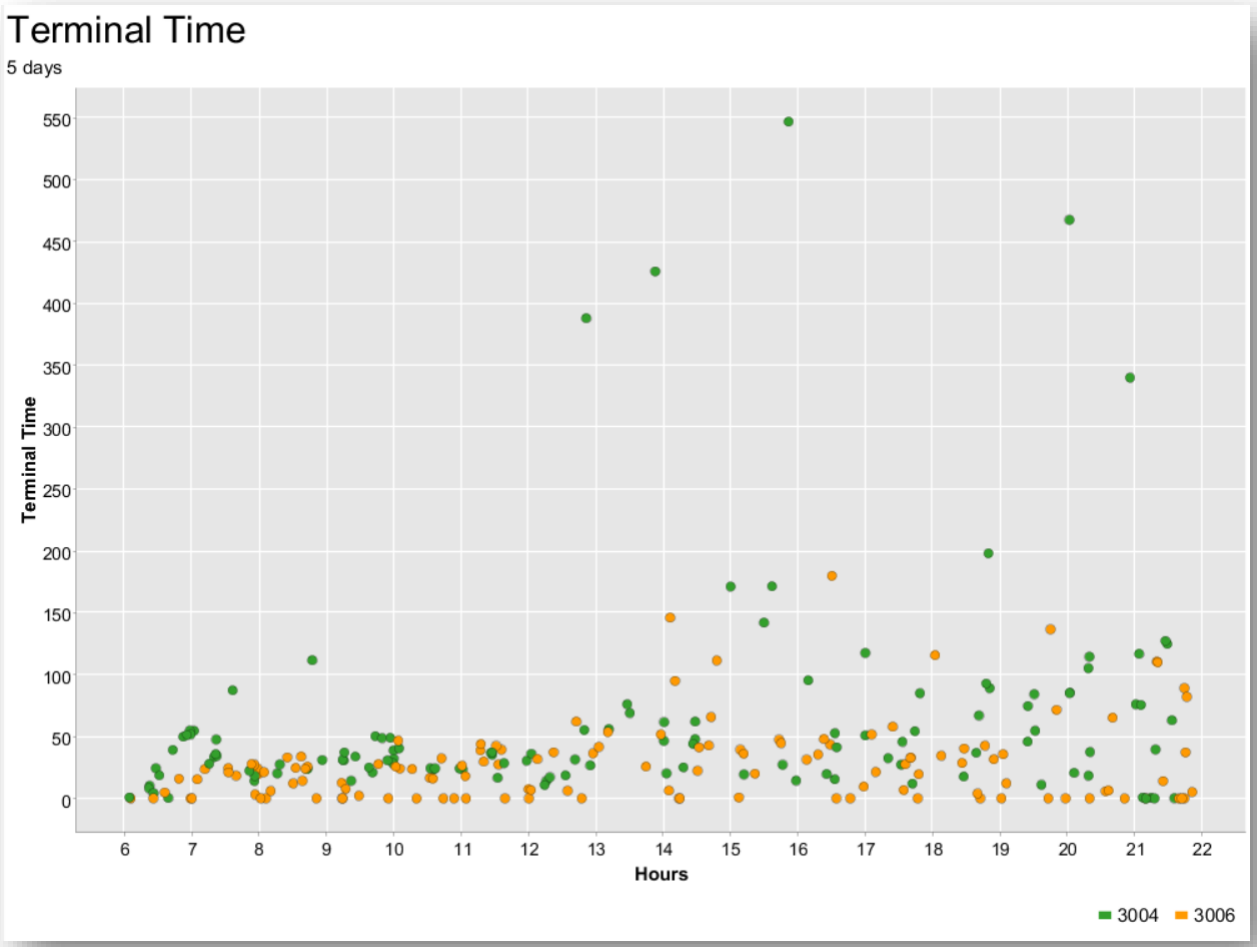
Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
RunningTime	19.967	79.933	37.645	9.601	92.185	2.013	5.268	20 80

Figure 29 Two-way running time 3004 to 3006 and backward

5.2.6.3 TERMINAL TIME

Terminal Time is the time that the bus stays in the terminal from arrival to the terminal until departure to the next trip. Figure 30 show the terminal time of each bus and terminal time general descriptive statistics respectively for terminal 3004 and 3006. The average terminal time was 35.31 minutes in a range of 0 minutes to 171.6 minutes.

Appendix C shows a scatter plot for one day terminal time. The scatter plots shows that in terminal 3004 the mean terminal time was higher than the terminal 3006, with 47.03 min and 20.52 min respectively.



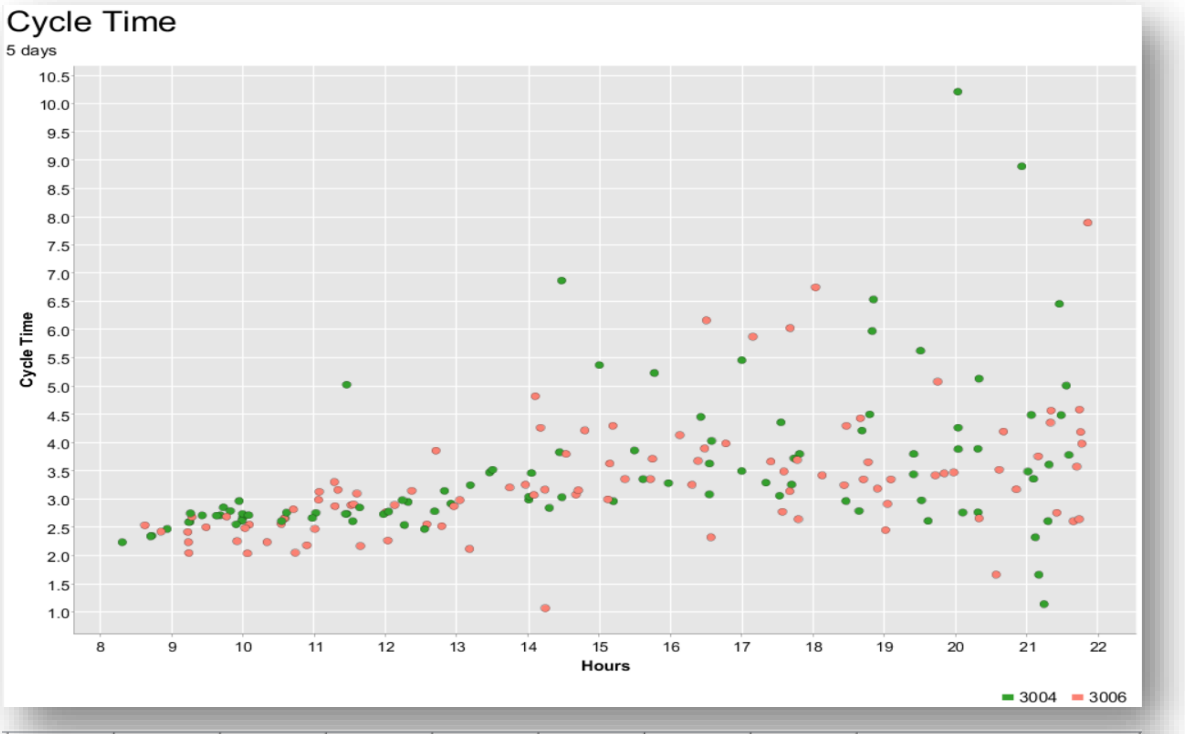
Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
Terminal Time	0	171.6	35.311	33.253	1,105.754	1.562	2.762	0 172

Figure 30 Terminal Time, Terminal 3004 and 3006

5.2.6.4 CYCLE TIME

The cycle time is the number of minutes needed to make a round trip on the route, including layover/recovery time as indicated in *TCRP 30* (Transportation Research Board, 1998). The cycle time scatter plot in Figure 31 represent the cycle time for each bus and cycle time descriptive statistics on Route 5 for five day of data.

The mean cycle time estimated is approximately 3.19 hr, with a range of 1.06 hr to 5.62 hr.



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
Cycle Time	1.066	5.627	3.194	0.795	0.632	0.644	0.665	1 6

Figure 31 Descriptive and Scatter Plot for Cycle on Terminal 3004

5.2.6.5 TRAVEL SPEED

The speed is a critical dynamic monitoring operational performance measure. Travel speed define the relationship between distance and travel time. The plots of the travel speed distribution and a box plot of the travel speed data are shown in

Figure 32.

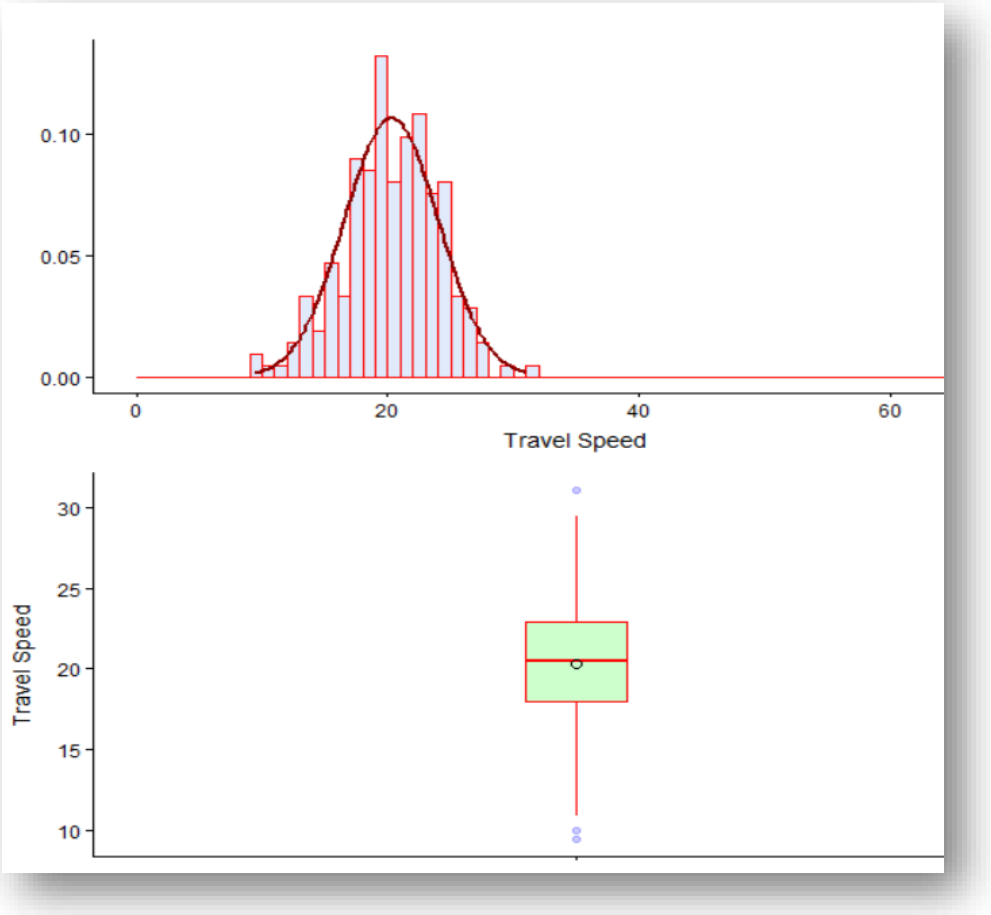
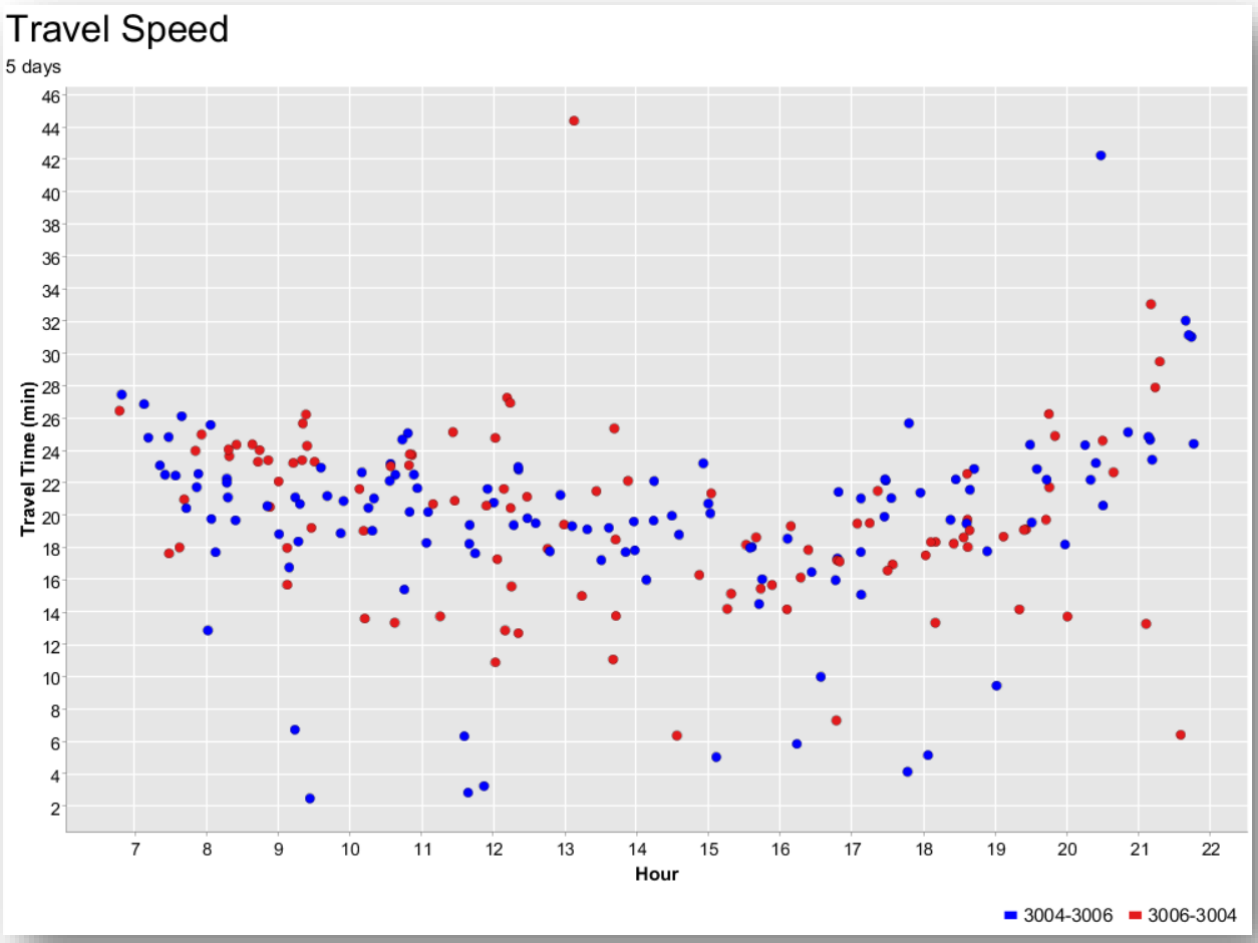


Figure 32 Histogram and Boxplot of the Mean Travel Speed of all buses from Route 5

The Figure 33 present the average travel speed for each bus on Route 5 by direction. The descriptive statistics in Figure 33 correspond to the mean speed of the buses on Route 5 from April 1 to April 5 of 2013. The figure shows a travel speed that ranges between 9.44 kph and 31.02 kph, with a mean of 20.27 kph.



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis
TravelSpeed	9.447	31.025	20.278	3.742	14	-0.239	0.137

Figure 33 Two-way travel speed 3004 to 3006 and descriptive statistics

5.2.6.6 RUNNING SPEED

The running speed is defined as the relationship between distance and running time. The descriptive statistics for running speed by bus, is shown in Table 5. The plots of the running speed distribution and a box plot of the running speed data are shown in Figure 34. Also, the Space-Time Diagrams are presented in Figure 35 and Figure 36.

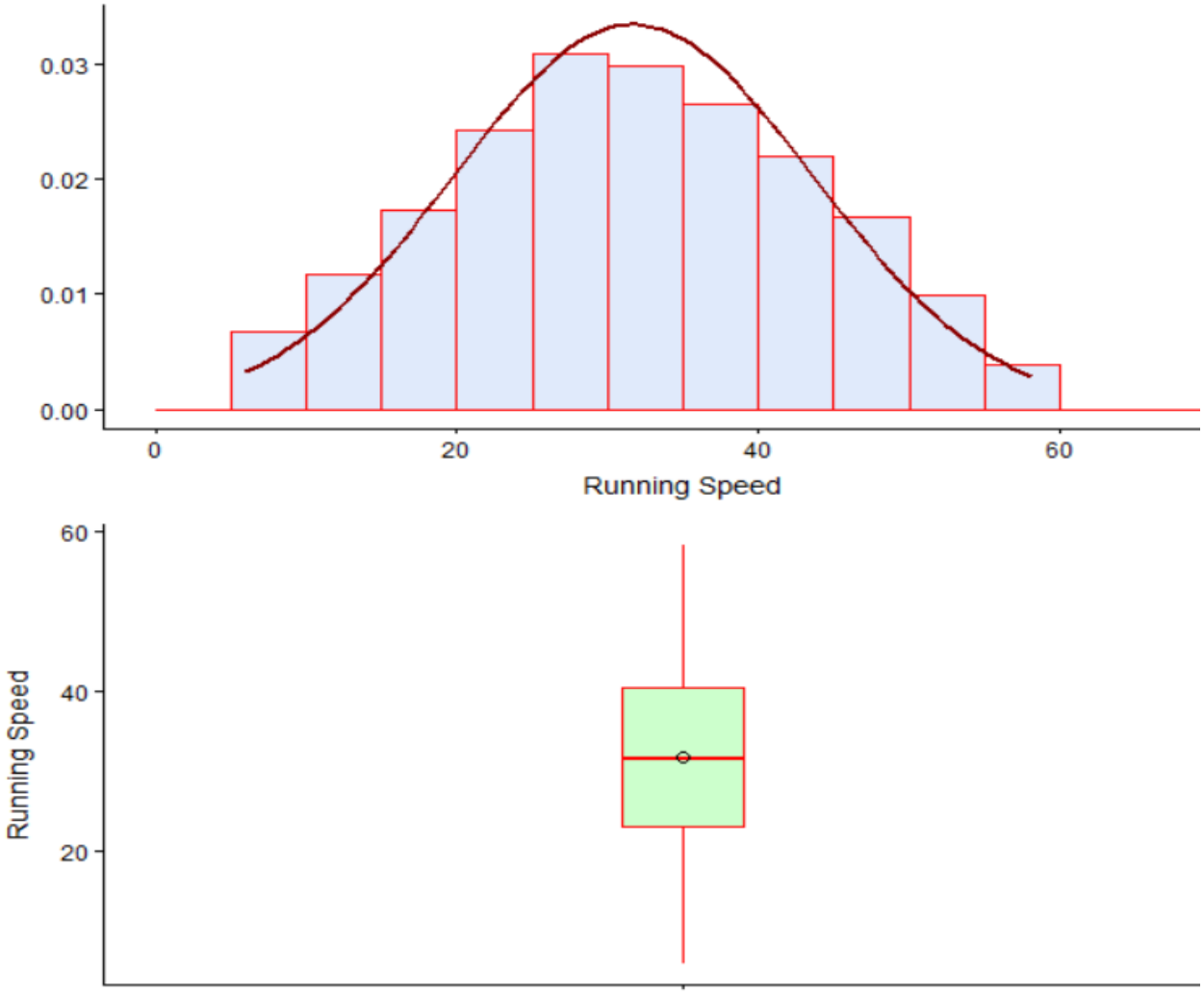


Figure 34 Histogram and Boxplot of the Mean Running Speed of All Buses from Route 5

Table 5 Descriptive Statistics of Running Speed by bus

nBus	Mean...	Standard deviation...	Min...	Max...	Kurtosis...	Skewness...	85.0-P^2 percentile...
826	29.754	12.679	3.047	66.732	-0.485	0.091	45.875
846	33.38	13.933	4.797	70.936	-0.441	0.14	48.203
861	31.912	13.184	1.878	60.968	-0.387	-0.034	45.828
862	32.655	13.555	3.176	73.167	-0.362	0.129	47.544
881	30.91	12.426	3.104	61.595	-0.493	-0.078	43.89
893	30.355	12.209	1.206	62.827	-0.352	0.188	44.724
927	31.225	12.732	0.61	64.462	-0.39	-0.049	44.127
993	29.861	11.676	1.603	67.983	-0.351	0.088	42.458
1003	34.798	12.861	8.29	75.904	-0.294	0.205	47.518
1006	32.098	12.751	1.896	69.516	-0.48	0.131	46.004
1007	34.091	13.856	1.409	76.938	-0.542	0.129	49.265
1018	31.573	13.083	0.968	87.218	-0.138	0.156	45.786
1020	32.984	13.52	1.087	79.329	-0.266	0.153	47.276
1021	31.34	12.298	2.494	61.087	-0.443	0.083	44.225
1023	33.445	12.998	0.421	66.773	-0.387	-0.004	45.359

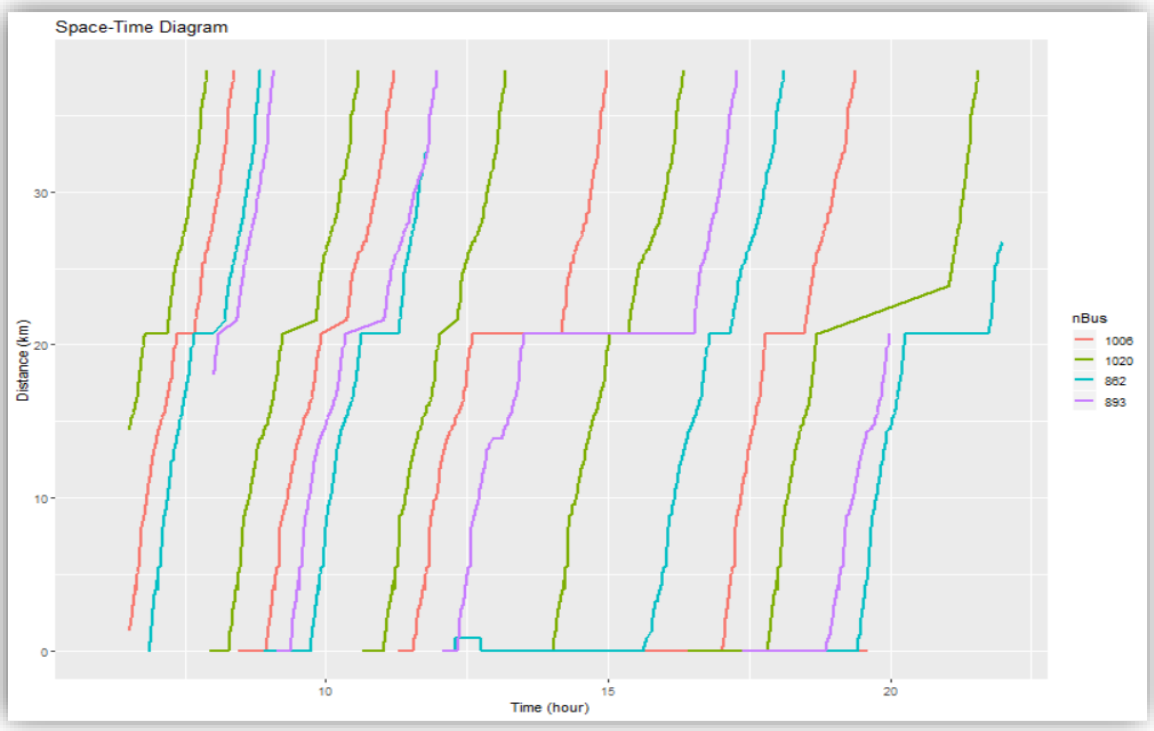


Figure 35 Space-Time Diagram for 4 Buses on Route 5

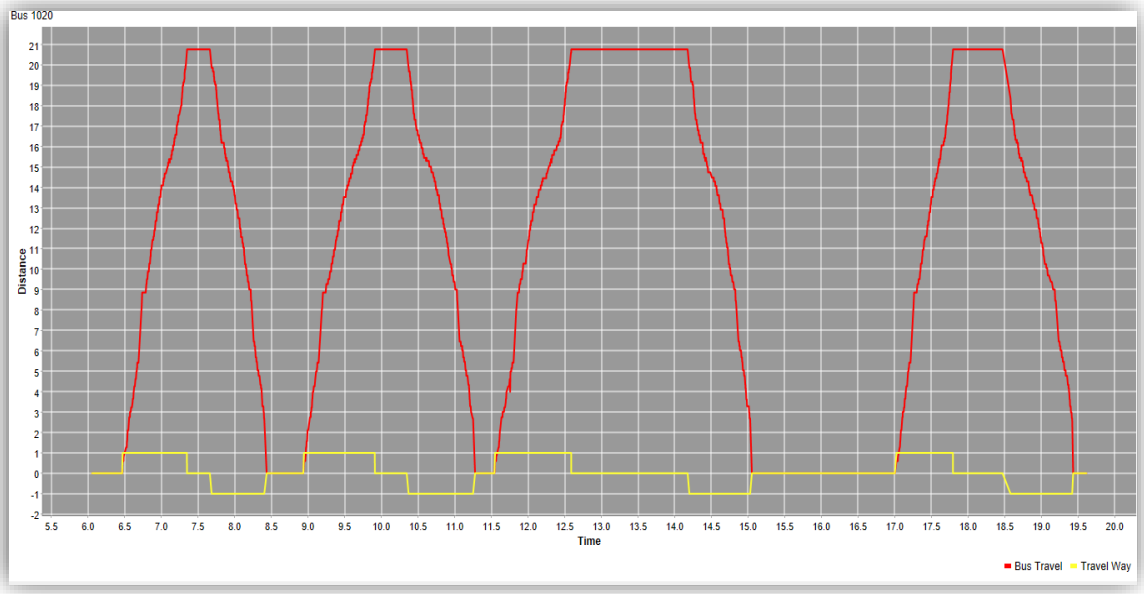


Figure 36 Space-Time Diagram Bus 1020

5.2.6.7 HEADWAY, REGULARITY INDEX AND COEFFICIENT OF VARIATION

The regularity index and coefficient of variation of headways are plotted in Figure 37. A sample of these metrics from Terminal 3004 (Iturregui) to bus stop 235 is presented in Table 6. The tendency of the coefficient of variation clearly follows the trend of the parameters of regularity, although with different magnitudes.

Table 6 Headway Regularity Index and Coefficient of Mean Variation for Bus Stops from 3004 to 235

oStop	nStop	Regularity Index	1-(Coefficient of Mean Variation)
1	3004	0.681	0.413
2	221	0.658	0.378
3	222	0.697	0.445
4	223	0.705	0.461
5	224	0.705	0.461
6	225	0.704	0.46
7	226	0.703	0.458
8	227	0.702	0.455
9	228	0.7	0.452
10	229	0.698	0.448
11	230	0.699	0.447
12	231	0.675	0.405
13	232	0.696	0.441
14	233	0.693	0.415
15	234	0.723	0.487
16	235	0.722	0.486

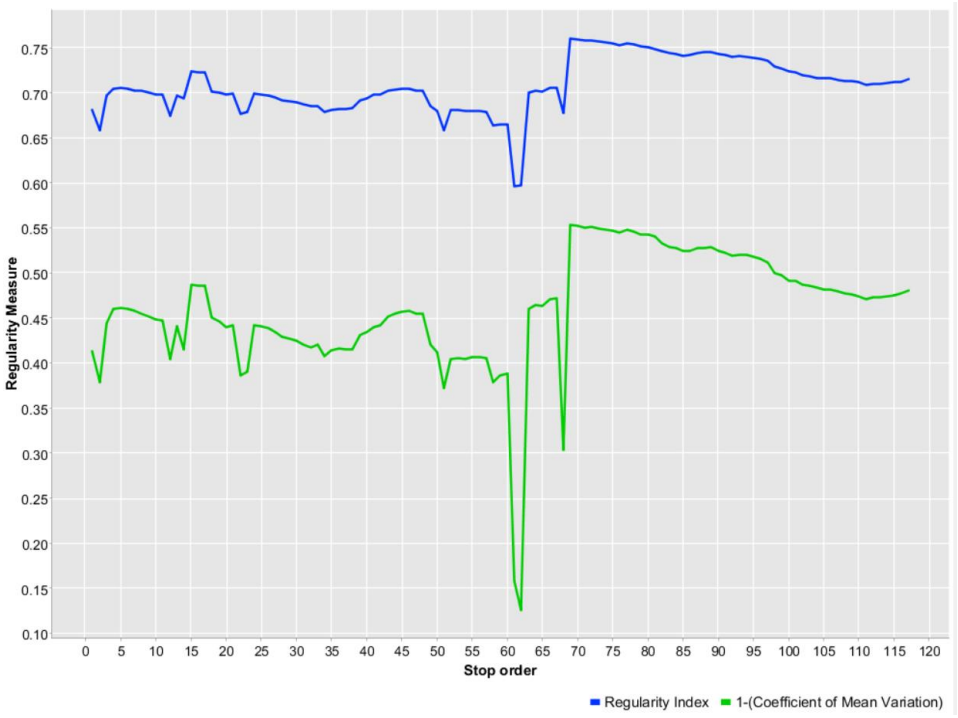


Figure 37 Line Plot of Regularity Index and Coefficient of Mean Variation

The magnitude 1-Coefficient of Mean Variation was estimated, to be congruent with the regularity index meaning

The regularity index and the coefficient of variation were estimated using the following equations, presented in the Transportation Research Record (Henderson, et al., 1991)

5.2.6.8 REGULARITY INDEX (R).

$$R = 1 - \frac{2\Sigma(h_r - H)r}{n^2 H}$$

Where:

h_r = series of headways.

r = 1,…, n, the rank of headways from smallest to largest

H = mean headway

5.2.6.9 COEFFICIENT OF MEAN VARIATION (CMV).

$$CMV = \frac{S}{\mathrm{H}}$$

Where:

S = the standard deviation of headways

H = mean headway

Other performance measures, estimated for Route 5 were: The Kilometers Traveled per Bus, the Passenger per Kilometer Index, the Cost per Kilometer, the Compliance Frequency Index and the Observed Schedule. These measures were calculated using the definitions detailed below. The results of the estimation for these measures are shown in Table 8 and Table 9. The Observed Schedule is shown in Table 7.

- **Kilometers Traveled per Bus**: Total amount of kilometers traveled by all buses divided by the number of buses.
- **Passenger per Kilometer Index:** Total number of passengers transported by buses, divided by the total amount of kilometers traveled by all buses.
- **Cost per Kilometer:** Passenger index per kilometer, multiplied by the fare. For estimation purposes, a fare of 0.75 cents per passenger was assumed.
- **Compliance Frequency Index:** Mean observed frequency divided by commercial schedule frequency. For estimation purposes, schedule frequency of 2 buses per hour was assumed.

Table 7 Extract of Observed Schedule for Terminal 3004 and 3006

nStop	nBus	Time
3004	1007	06:47:02
3004	862	06:52:20
3004	1007	07:19:43
3004	1018	07:36:22
3004	1020	08:16:09
3004	1006	08:26:15
3004	862	08:54:09
3004	1006	08:56:06
3004	893	09:21:57
3004	1007	09:29:20
3004	862	09:43:33
3004	1007	09:59:54
3004	1020	11:01:20
3004	1006	11:32:30
3004	893	12:03:24
3004	862	12:15:52
3004	893	12:18:56
3004	1007	12:29:05
3004	862	12:46:23
3004	1007	12:55:04
3004	1020	13:14:59
3004	1018	13:40:56
3004	1020	14:00:38
3004	1018	14:28:23
3004	1023	14:34:55
3004	1006	15:03:12
3004	862	15:36:55
3004	1007	15:41:41
3004	1023	16:09:09
3004	1020	16:25:03
3004	1007	16:32:49
3004	1006	17:00:04
3004	1018	17:18:36
3004	893	17:21:54
3004	1018	17:31:51
3004	1020	17:48:37
3004	862	18:11:28
3004	1023	18:37:01
3004	893	18:50:53
3004	862	19:24:51
3004	1006	19:36:52
3004	1018	19:46:28
3004	1023	20:02:19
3004	1007	20:30:07
3004	1018	21:01:04
3004	1007	21:33:15
3004	862	22:27:30

nStop	nBus	Time
3006	1020	06:47:43
3006	1023	06:58:50
3006	1020	07:11:37
3006	1006	07:21:18
3006	862	07:39:31
3006	1006	07:39:37
3006	862	07:59:05
3006	893	08:05:57
3006	1007	08:30:07
3006	1020	09:14:20
3006	1023	09:28:55
3006	1006	09:54:57
3006	893	10:20:21
3006	862	10:38:14
3006	1007	10:56:32
3006	862	11:17:04
3006	1007	11:35:52
3006	1023	11:39:01
3006	1018	11:59:57
3006	1020	12:00:24
3006	1006	13:09:28
3006	893	13:30:28
3006	1007	13:57:38
3006	1006	14:01:23
3006	893	14:04:06
3006	1006	14:10:28
3006	893	14:24:21
3006	1007	14:40:29
3006	1020	15:01:37
3006	893	15:09:08
3006	1020	15:21:40
3006	1018	16:07:53
3006	893	16:30:17
3006	862	16:48:04
3006	1023	17:07:35
3006	862	17:09:28
3006	1007	17:28:31
3006	1023	17:40:36
3006	1006	17:47:57
3006	1018	18:27:46
3006	1006	18:28:08
3006	1020	18:42:26
3006	1018	19:02:51
3006	1007	19:45:05
3006	893	19:58:38
3006	862	20:15:25
3006	1023	20:51:07

Table 8 Estimation of Km per bus

Km per Bus	131.041
Sum(Load)	29,562
Passenger Km Index	32.228
Cost per Km	24.171

Table 9 Estimation of Average Headway, Average Frequency and Compliance Frequency Index

nParada	Average Headway	Average Frecuency	Compliance Frequency Index
3004	21.387	2.806	1.403
3006	21.919	2.737	1.369

5.2.7 MODELING

The fourth part of the process is linked to the nature of the response variable. The boarding of passengers by stop is the response variable according to the variables mentioned in The variables selected for the analysis are: Average Income per capita measured in thousands of dollars per year, and Mid distance between two consecutive stops measured in meters. The variables hospitals in the area, schools in the area, governmental offices in the area, industries in the area, recreation in the area and tourism zones in the area, are dichotomous variables measured as 1 or 0 representing the presence (1) o absence (0) of this type of travel generators in the zone. These variables were grouped as shown in Table 2.

Table 2. The variable "boardings" is of an integer nature, with positive range, so it will be managed as a counting variable. Counting variables can be modeled using the Poisson or Negative Binomial model. The process starts with an analysis of the variability of the data through the estimation of its dispersion. Model choice will depend on the value of the dispersion.

The last part of the process involves estimating several nested models and choosing the best one as a function of goodness-of-fit parameters and information criteria such as AIC or BIC.

The entire preparation, processing, and modeling process was made as a sub-process of the CRISP-DM methodology shown in Figure 21. This sub-process, which begins with the use of Knime Analytics for mining and ends with the integration of Knime and R-Studio, for the statistical modeling of the data, is shown in Figure 27 .

The modeling process was divided in two stages, to explain the relationship that occurs in the stops, as shown in Figure 38:

• Construction of performance measures that can be monitored in real-time.

• The development of a statistical model that links these measures with the passenger boarding and the activity system that takes place around the stops.

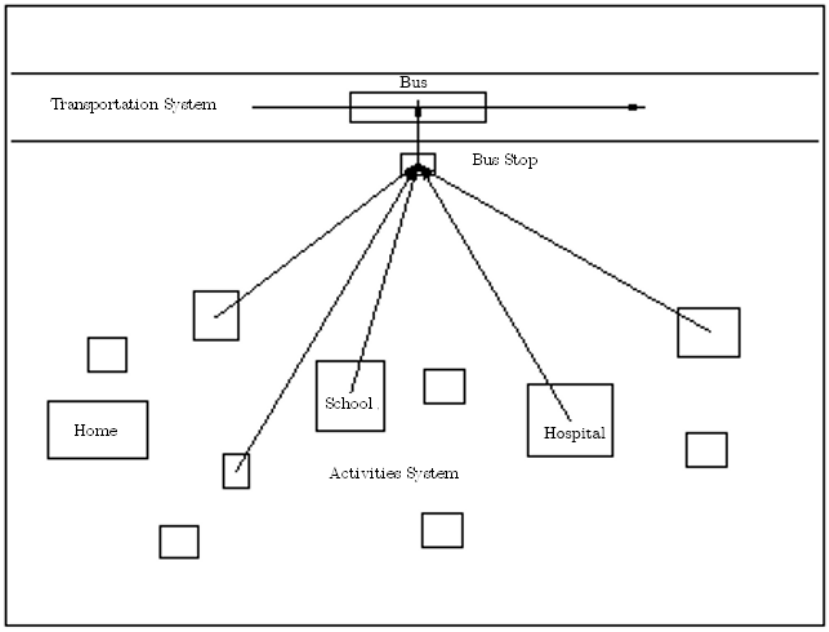


Figure 38 Relationship and Exchange of the Transportation System and the Activities System in a Stop

5.2.8 EVALUATION AND IMPLEMENTATION

This is the last part of the process and will be executed in coordination with the responsible agency. This step involves evaluating the recommendations made by the analysis and developing policies aimed at improving the system through implementation. This part of the process is not part of this study.

5.2.9 RESULTS

The results were grouped into two groups, according to the stages in which the modeling process was divided: performance measures of the transportation system and statistical model of the activity system.

6 PROPOSED PERFORMANCE MEASURES

The proposed performance measures presented in this chapter, aim to evaluate the quality of the service considering parameters that are related to ridership. Frequency and adherence are two of the metrics that significantly affect boarding. Frequency serves as a measure of regularity, and adherence, as a measure of the degree of compliance with itineraries according to the established commercial headway. Another performance measure that is directly correlated with boarding is the number of buses that stopped on a stop. The proposed measures that follow were divided considering the headway and regularity, the number of stops traveled for each bus, and the number of buses stopped on a stop.

6.1 BASED ON HEADWAY AND REGULARITY.

Four performance measures are proposed to determine the quality of service of the transportation system based on headways and regularity. These four measures compare the actual headway calculated using AVL/GPS data to the commercial headway established by the agency in charge of operating the system. The performance measures to be estimated will have the functional form shown in the following equations:

$$HS = f(H_o, H_c, t) \quad (1)$$

$$HPF = f(H_o, H_{max}, n) \quad (2)$$

$$IHS = f(HS, HPF) \quad (3)$$

$$\text{MVR} = f(H_o, n, H) \quad (4)$$

Where:

HS: Headway Score. Defined as the number of times a bus moves away from the commercial headway

HPF**:** Headway Peak Factor. Defined as a factor that measures how homogeneous the observed headway is

IHS**:** Inflated Headway Score. Defined as the inflated score (IS), when affected by the Headway Peak Factor (*HPF)*

MVR: Mean Variation Rate

***H_o*:** Observed Headway. Defined as the time between a bus and the next one, observed at the stop

***H_{max}*:** Maximum Headway. Defined as the maximum observed headway during a particular period of time time between a bus and the next one, observed at the stop

***H_c*:** Commercial Headway. Defined as the business time, assigned by the company, in which a bus must go out regularly to make its tour

H: Mean Headway

***t*:** Headway tolerance

***n*:** Number of headways observed in the analysis period.

6.1.1 HEADWAY SCORE (HS)

The Headway Score is defined as the ratio between the headway deviation from the commercial headway (average observed bus headway (***H_o***), relative to the commercial headway (***H_c***)), measured at each stop or terminal, and the headway tolerance (***t***), established by the agency or system operating entity. The tolerance may vary among agencies; however, if this metric is to be used to compare performance between agencies, the tolerance should be established in advance. For comparison purposes, 20% of the headway is recommended as a value for the variable tolerance. However, a variable value depending on the size of the commercial headway could also be considered to set the value of tolerance.

The Headway Score (HS) is one of the variables depicted in

Figure 40. The value estimated with this metric indicates the deviation from the commercial headway, measured in "tolerance" units.

$$HS = \frac{H_o - H_c}{t} \qquad (5)$$

6.1.2 HEADWAY PEAK FACTOR (HPF)

The Headway Peak Factor is defined as the ratio between the sum of the headways in a stop or route ($\sum H_o$), in a specified period, and n-times the maximum headway recorded in that period (*nHmax*), where *n* is the number of observed headways in the period. The estimated values with this measure indicate the presence and severity of very long headways of the buses at the stop where this metric is calculated. This metric also show the variablility of the observed headways with respect to the maximum headway. This number can vary between 0 and 1. Values close to the unit reflect uniform headways; on the other hand, the farther these values are from one (1), the highest the presence of very long headways.

$$HPF = \frac{\Sigma H_0}{nH_{max}} \qquad (6)$$

6.1.3 MEAN VARIATION RATE (MVR)

The Mean Headway Variation Rate is defined as the percentage of variation of the observed headway with respect to the mean headway.

$$MVR = 1 - \frac{1}{n}\Sigma\frac{|H_o - \mathrm{H}|}{\mathrm{H}} \qquad (7)$$

A plot for the estimation of the HPF and MVR joined with other measures of regularity, including the regularity index and the coefficient of variation is shown in Figure 39. Table 10 shows a sample of the type of data included in Figure 39 . The tendency observed on the figures corresponding to the variables HPF and MVR is like the trend of the variables Regularity and Coefficient of Mean Variation (CMV). To make it easy to compare, the magnitudes depicted for the variable CMV in

Figure 39 Comparison of Headway Peak Factor and Mean Variation Rate with other regularity performance measures correspond to the calculation of one minus the Coefficient of Mean Variation (1-CMV)

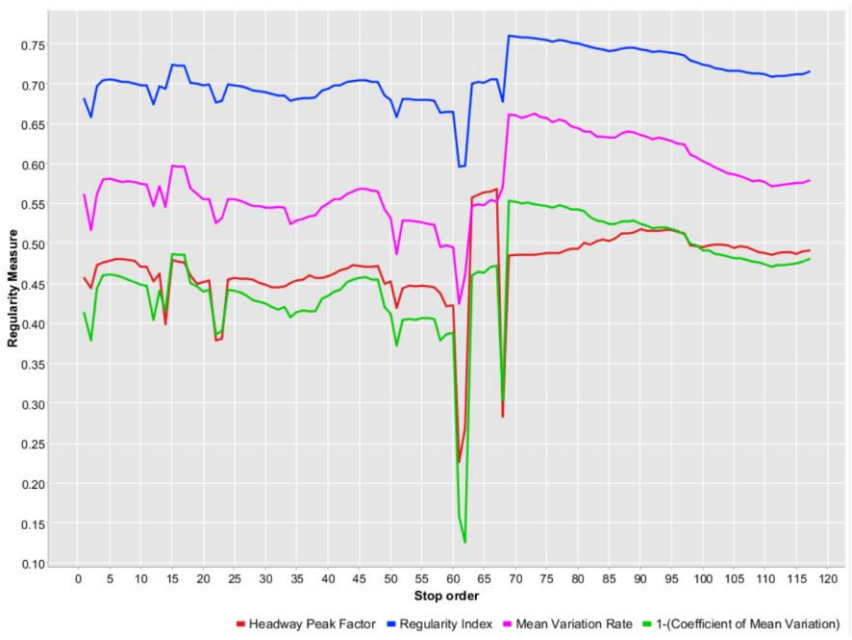


Figure 39 Comparison of Headway Peak Factor and Mean Variation Rate with other regularity performance measures

The magnitude 1-Coefficient of Mean Variation was estimated, to be congruent with the regularity index meaning

6.1.4 INFLATED HEADWAY SCORE

*This parameter is defined as the ratio between the headway score (*HS*) and the Headway Peak Factor (HPF). This value measures the combined effect between the arrival time deviation of the buses at a stop, relative to the expected headway, and the uniformity to which the buses arrive.*

Figure 40 shows the comparison between HS and IHS. The inflated value is always higher or equal to the HS. It is only equal when HPF is one meaning when there is uniformity in the headways. Therefore, IHS represents a measure that not only indicates if the headways are longer than expected but also if there is uncertainty in the length of the headway. IHS could be used as a measure of the uncertainty and annoyance produced in passengers, for not knowing exactly when a bus arrived, nor the time they should wait for it.

$$IHS = \frac{HS}{HPF} \qquad (8)$$

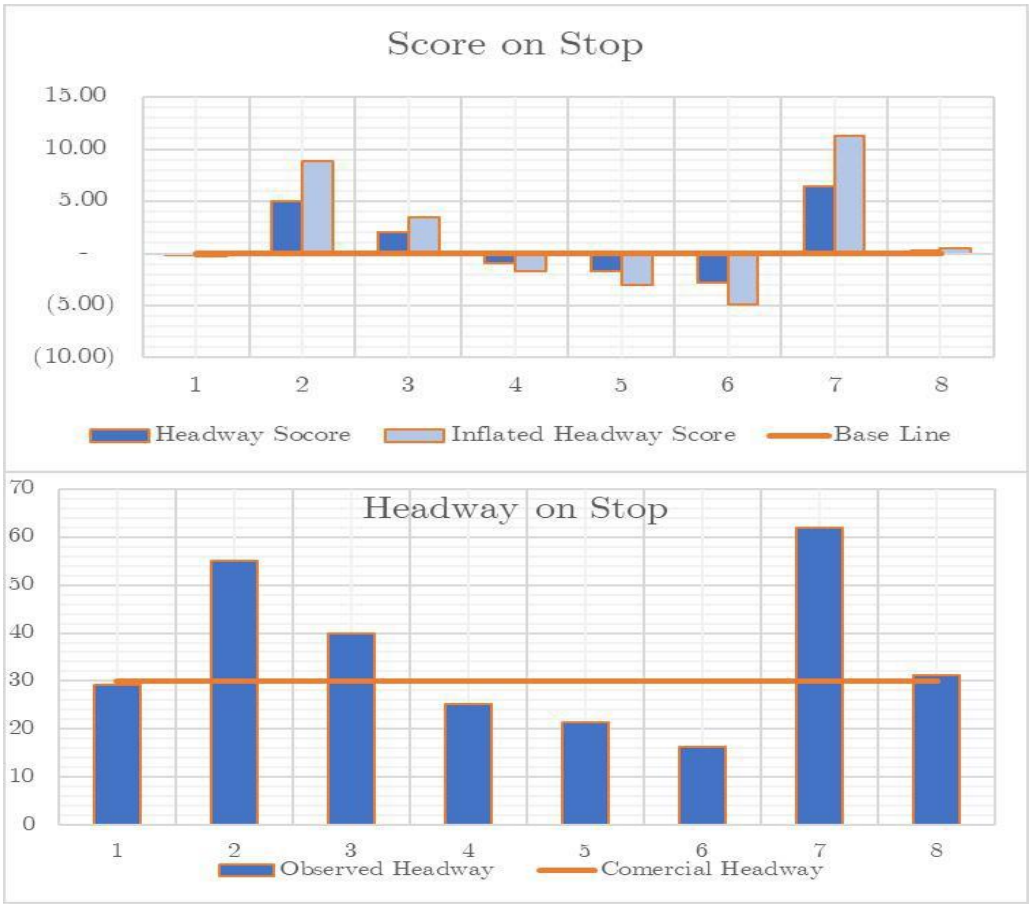


Figure 40 Comparison between the Observed Headway Measured at a Stop, the Commercial Headway and the Headway Affected by the Uniformity of Arrivals or Inflated Headway

The new proposed metrics were calculated in all the bus stops on Route 5. A sample of the values calculated for these parameters is presented in Table 10 corresponding to the calculations from terminal 3004 to bus stop 264. It can be observed that there are differences between the Regularity Index and the Headway Peak factor and Mean Variation Rate, but that all of them follow a similar trend.

Table 10 New Performance Measure Calculated in Bus Stop, from Terminal 3004 to Bus Stop 264

oStop	nStop	Headway	Regularity Index	Mean Variation Rate	Headway Peak Factor
1	3004	21.387	0.679	0.526	0.348
2	221	36.257	0.687	0.539	0.42
3	222	37.615	0.71	0.565	0.435
4	223	36.302	0.728	0.6	0.498
5	224	36.312	0.728	0.601	0.5
6	225	36.291	0.727	0.599	0.502
7	226	36.277	0.726	0.597	0.503
8	227	36.269	0.725	0.596	0.502
9	228	36.249	0.723	0.594	0.499
10	229	36.242	0.721	0.591	0.492
11	230	36.238	0.721	0.59	0.491
12	231	34.865	0.692	0.556	0.486
13	232	36.247	0.719	0.588	0.484
14	233	39.517	0.718	0.563	0.42
15	234	37.815	0.749	0.62	0.504
16	235	37.814	0.748	0.619	0.501
17	236	39.775	0.733	0.592	0.466
18	237	36.528	0.729	0.598	0.485
19	238	36.659	0.731	0.597	0.476
20	239	36.715	0.73	0.59	0.479
21	240	36.752	0.732	0.593	0.482
22	241	38.437	0.713	0.566	0.405
23	242	36.918	0.74	0.602	0.484
24	243	36.982	0.741	0.603	0.487
25	244	37.01	0.742	0.604	0.489
26	245	37.685	0.742	0.607	0.497
27	246	37.683	0.741	0.601	0.496
28	247	37.669	0.738	0.597	0.495
29	248	37.665	0.738	0.597	0.491
30	249	37.646	0.739	0.597	0.489
31	250	37.63	0.738	0.598	0.486
32	251	37.641	0.738	0.599	0.486
33	252	37.604	0.732	0.597	0.485
34	253	37.623	0.723	0.571	0.487
35	254	37.573	0.723	0.573	0.49
36	255	37.591	0.723	0.573	0.491
37	256	37.401	0.725	0.584	0.493
38	257	37.385	0.723	0.583	0.488
39	258	37.357	0.727	0.593	0.487
40	259	37.34	0.728	0.596	0.488
41	260	37.335	0.729	0.596	0.49
42	261	37.32	0.729	0.596	0.494
43	262	37.293	0.732	0.601	0.496
44	263	37.271	0.733	0.603	0.5
45	264	37.289	0.733	0.605	0.498

6.1.5 LEVEL OF SERVICE OF HEADWAY

This parameter uses the headway type of variables previously defined to measure the quality of the service offered from the agency point of view. Any of the headway regularity variables could be used to make the graphic that will be used to define the level of service. The Regularity index (R), The Headway Peak Factor (HPF) and the Headway Score (HS) will be used here to show how this level of service of headway can be defined and used. Six (6) levels of service ranging from the

most favorable (*A*) to the most unfavorable (*F*) will be determined. The suggested thresholds for each of the levels of service proposed are shown in Table 11, and Table 12. Figure 41 shows a graphical representation of these thresholds and allow for the combination of variables to define the level of service.

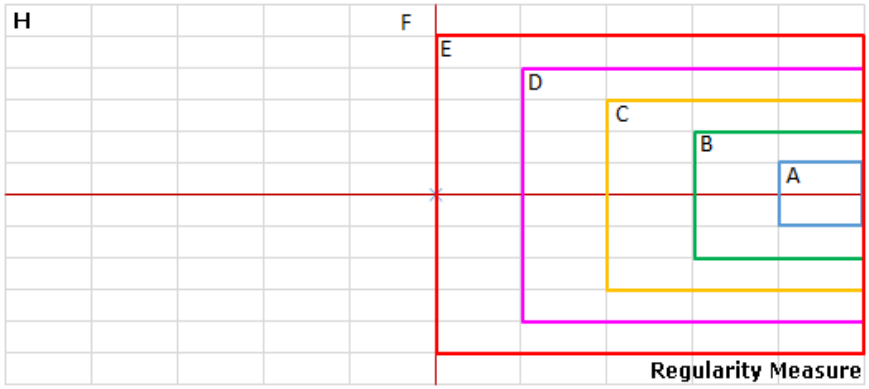


Figure 41 Level of Service for the System from the Point of View of Adherence

The values presented in Table 11 and the colored rectangles representing the level of service limits in Figure 41 are proposed to depict the level of service. As the value of the uniformity factor moves away from the unit (bus headways are less uniform) the level of service decreases. At the same time, as the headway scores move away from zero (buses headways are higher or lower than the commercial headway), the service also decreases.

Table 11 Level of Service for the System from the Point of View of Adherence Measure with Headway Peak Factor

LOS	HPF		HS	
A	0.9	1	-0.5	0.5
B	0.8	0.9	-1	1
C	0.7	0.8	-1.5	1.5
D	0.6	0.7	-2	2
E	0.5	0.6	-2.5	2.5
F	> 0.6		< -2.5	> 2.5

Table 11presents the proposed level of service thresholds considering the variable Headway Peak Factor to represent the headway regularity. Table 12 presents the equivalent level of service thresholds using the variable Regularity Index. Both Tables use the variable Headway Score to complement the regularity variable in the definition of the level of service.

Table 12 Level of Service for the System from the Point of View of Adherence Measure with Regularity Index

LOS	R		HS	
A	0.9	1	-0.5	0.5
B	0.8	0.9	-1	1
C	0.7	0.8	-1.5	1.5
D	0.6	0.7	-2	2
E	0.5	0.6	-2.5	2.5
F	> 0.6		< -2.5	> 2.5

A plot of the headway scores and the headway peak factors of the 117 stops on AMA Route 5, chosen for the analysis, is shown in Figure 42. In this graph, the level of service of the route, with respect to the adherence of the buses to the commercial headway, is between *E* and *F* for the data considered in these calculations.

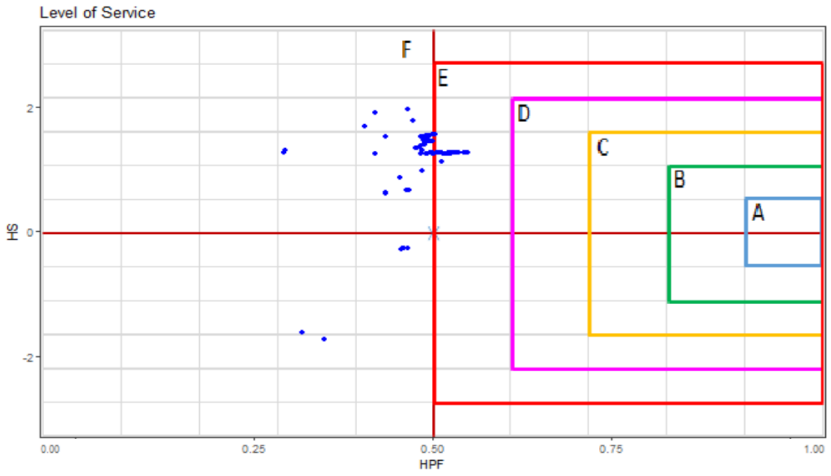


Figure 42 Level of Service for the AMA from the HPF Perspective

Figure 43 presents a similar plot but in this case using the Headway Regularity Index as a regularity measure. This plot shows that in this case, considering the commercial headway adherence, the route's service level is between *C* and *D*. The Regularity Index is one of the existing performance metrics. The level of service associated with the regularity index makes the route look better than the proposed Headway Peak Factor. The Regularity Index smooths out the effects of the peaks in the data. Therefore, the proposed measure is a better indication for situations that can be corrected in real-time if an AVL system is available. Calculating the level of service using R tends to be more conservative that calculating it with HPF.

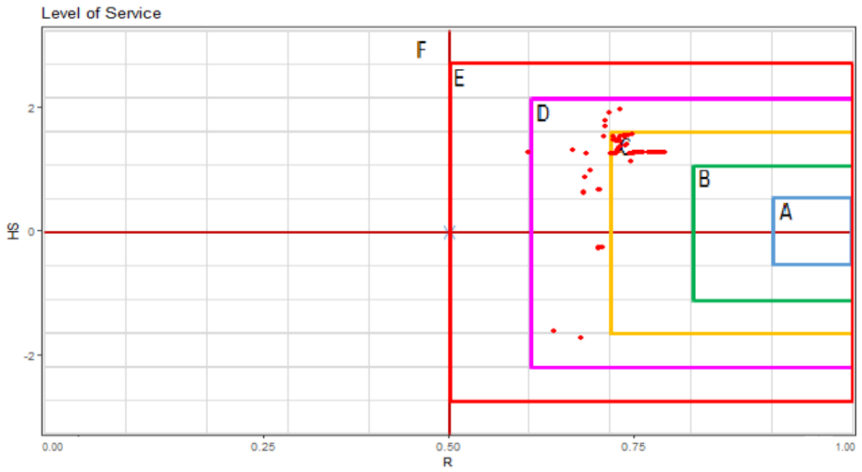


Figure 43 Level of Service for the AMA from the R Perspective

6.2 Based on Number of Stop Traveled for a Bus, and Number of Buses Stopped on a Stop.

To take into consideration the performance of the system from the system point of view, five additional performance measures are proposed. These five measures look at buses traveling by a series of stops during a time period and buses passing by each stop during a specific time period. The performance measures to be estimated will have the functional form shown in the following equations:

$$Ksr = f(Ns, Lr) \quad (9)$$

$$STS = f(Krs, V) \quad (10)$$

$$TSI = f(Nst, STS) \quad (11)$$

$$SBS = f(Nh, f) \quad (12)$$

$$SBI = f(No, SBS) \quad (13)$$

Where:

Ksr: Density of Stop per Route. Defined as the number of stops per km in a route of length *Lr*.

***SST*:** Schedule Stop Traveled. Defined as the number of scheduled stops where a bus should stop in an hour.

***TSI*:** Traveled Stop Index. Defined as the percent of stops traveled, considering SST as reference.

***SBS*:** Scheduled Bus Stopped. Defined as the number of buses scheduled, that should pass by a stop in an hour.

***SBI*:** Stopped Bus Index. Defined as the percent of buses stopped, considering SBS as reference.

***Ns*:** Number of Stops of the route.

Lr: Length of a Route.

***V*:** Mean Travel Speed of the buses on the route.

***Nst*:** Number of stops traveled by a bus in an hour.

***Nh*:** Number of hours of the period of analysis.

***f*:** frequency of the buses.

***No*:** Number of buses observed in a stop in an hour.

6.2.1 STOP DENSITY PER ROUTE

This parameter is defined as the ratio between number of stops in a route and the length of the route. The stop density per route, is a magnitude that express the number of stop per unit of length in a route.

$$Ksr = \frac{Ns}{Lr} \qquad (14)$$

6.2.2 SCHEDULE STOP TRAVELED

This parameter is defined as the product of the Stop Density per Route (Ksr) and the travel mean speed of the buses on that route. The schedule stop traveled, measure the number of stops that a bus can pass in an hour, when traveling at the mean travel speed.

$$SST = Krs\ x\ V\ \left(\frac{stop}{h}\right) \qquad (15)$$

6.2.3 TRAVELED STOP INDEX

This parameter is defined as the ratio between the number of stops traveled by a bus in a route and SST. This parameter compares the number of stops passed and the schedule stop to be traveled.

$$TSI = \frac{Nst}{SST} \qquad (16)$$

6.2.4 SCHEDULED BUS STOPPED

This parameter is defined as the product of the number of hours taken for analysis and the frequency of buses of the route. This parameter measures the number of buses that is supposed to pass through a stop during the analysis period.

$$SBS = N_h \, x \, f \qquad (17)$$

6.2.5 STOPPED BUS INDEX

This parameter is defined as the ratio between number of buses stopped on a stop of a route and SBS. This parameter compares the number of buses stopped in a stop of a route and the scheduled number of buses that should pass by each stop.

$$SBI = \frac{N_O}{SBS} \qquad (18)$$

The index of stopped buses can be closely linked to the boarding level. From the point of view of the buses, a low index can indicate that:

- The buses are not stopping at the stop because they are full
- The buses are not stopping at the stop because there are no passengers at this stop.

From the point of view of the passengers, a low level of boarding can indicate that:

- The passengers go to another stop because the buses do not stop at this stop.
- There are few passengers in the influence area of this stop.

Table 13 shows the estimates corresponding to the metrics SBS, SST and Krs from Route 5. The values indicate that at an average travel speed, 31 buses should pass by each stop of this route during one day of service, every bus should pass by 60 stops every hour, and that in average, there are 3 stops per km along this route.

Table 13 Stop Density per route (Krs), Scheduled Stop Traveled (SST) and Scheduled Bus Stopped (SBS) on Route 5

SBS	SST	Krs
31.598	59.376	2.969

Table 14 shows a sample of the estimates corresponding to SBI on Route 5. The data shown corresponds to the segment between Terminal 3004 (Iturregui) and Stop 250. For each stop, this table present the values for *No* and *SBI*.

shows the estimates corresponding to SBI on Route 5. The data in Table 14 are plotted in. The stops that have a low index correspond to the stops where buses stop less.

Table 14 View of Stopped Bus Index (SBI) for stops on Route 5

S nParada	I No	D SBI
3004	29	0.918
221	29	0.918
222	27	0.854
223	26	0.823
224	26	0.823
225	26	0.823
226	26	0.823
227	26	0.823
228	26	0.823
229	26	0.823
230	26	0.823
231	27	0.854
232	26	0.823
233	24	0.76
234	25	0.791
235	25	0.791
236	25	0.791
237	26	0.823
238	26	0.823
239	26	0.823
240	26	0.823
241	25	0.791
242	25	0.791
243	26	0.823
244	26	0.823
245	26	0.823
246	26	0.823
247	26	0.823
248	26	0.823
249	26	0.823
250	26	0.823

The complete data set corresponding to the Stopped Bus Index for all the stops along Route 5 is plotted in Figure 44. The stops that have a low index correspond to the stops where buses stop less.

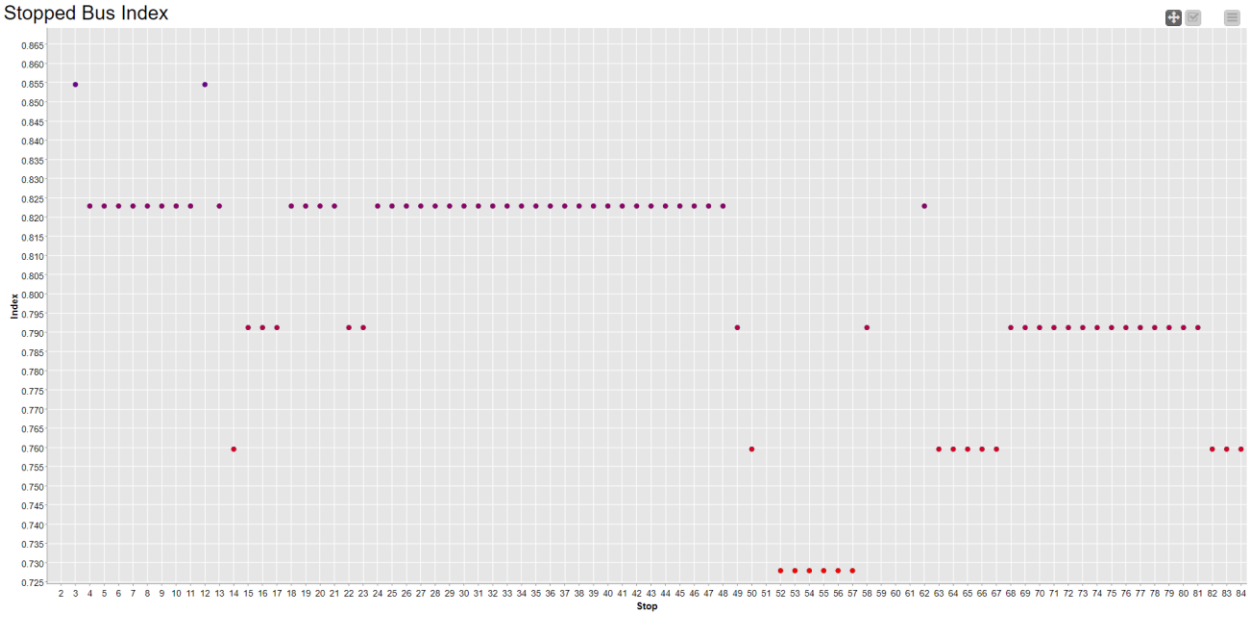


Figure 44 Fixed View of Dynamic Stopped Bus Index

Table 15 shows a sample of the estimates corresponding to TSI for each bus per hour, representing the percentage of stops traveled considering SST as a reference. This table presents the data for each bus including the bus number, the hour of reference, the value of Nst and the Travel Stop Index.

Table 15 Traveled Stop Index for bus on Route 5

nBus	Hour	Nst	TSI
862	6	15	0.253
1006	6	34	0.573
1007	6	53	0.893
1018	6	1	0.017
1020	6	57	0.96
1023	6	11	0.185
862	7	48	0.808
1006	7	52	0.876
1007	7	38	0.64
1018	7	1	0.017
1020	7	57	0.96
1023	7	6	0.101
862	8	56	0.943
893	8	55	0.926
1006	8	36	0.606
1007	8	42	0.707
1020	8	40	0.674
862	9	18	0.303
893	9	50	0.842
1006	9	56	0.943
1007	9	38	0.64
1020	9	38	0.64
1023	9	14	0.236
862	10	42	0.707
893	10	23	0.387
1006	10	40	0.674
1007	10	59	0.994
1020	10	39	0.657

Figure 45 shows the cumulative passed stop for each bus in all day. In this graph, buses 1018 and 1023 did not accumulate the expected number of stops. This is also observed in Table 15.

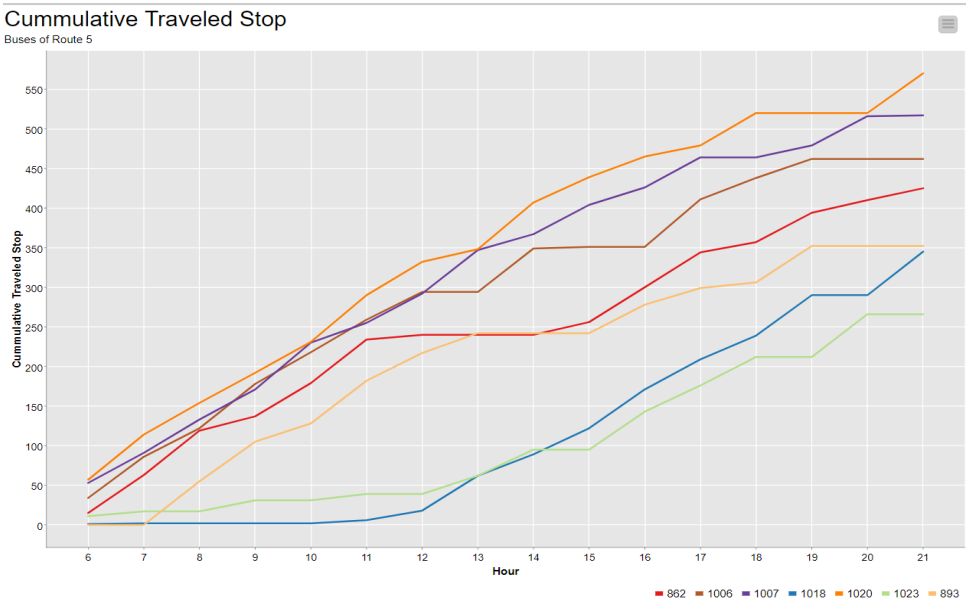


Figure 45 Cumulative Traveled Stop by bus on Route 5

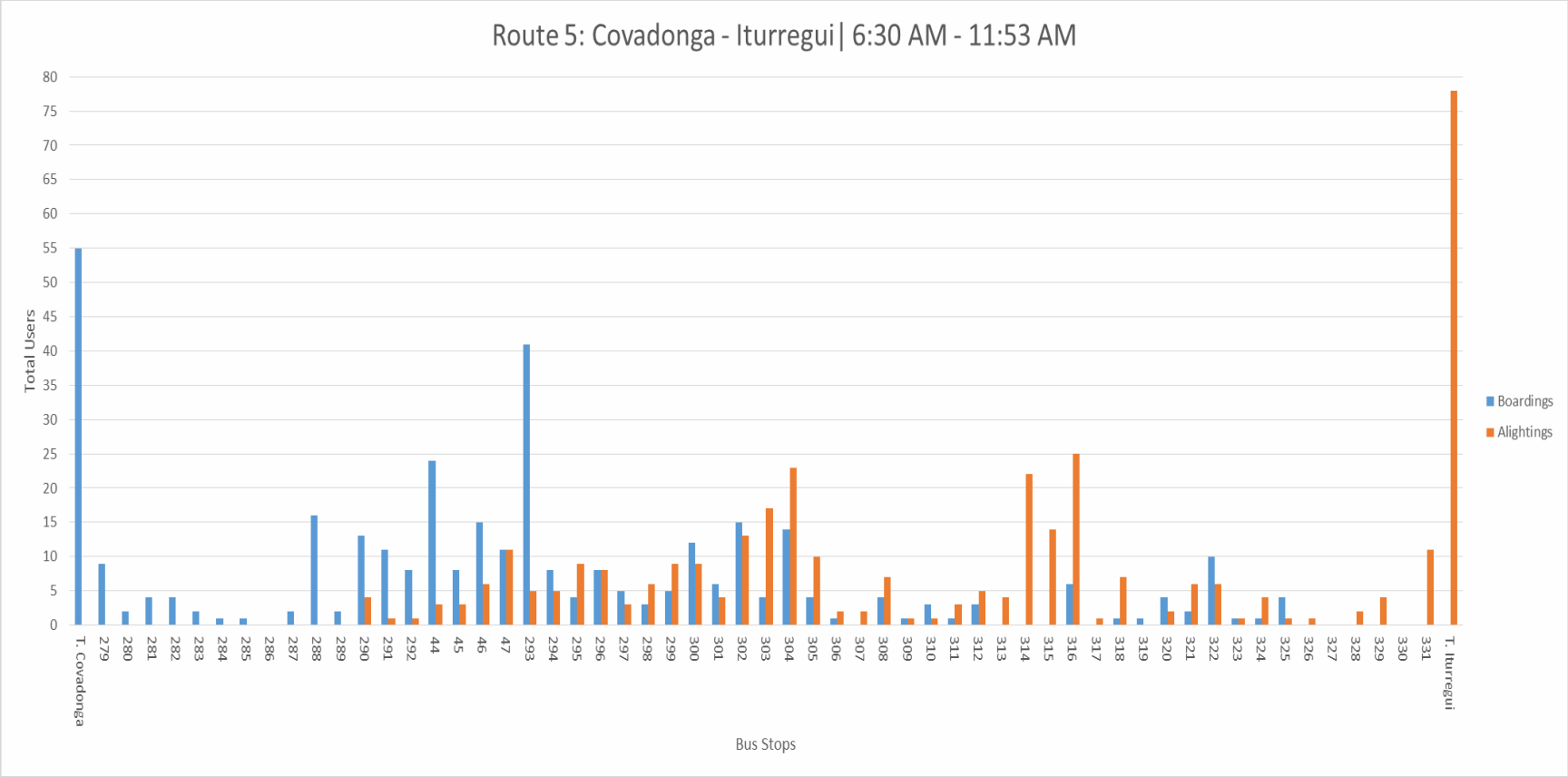
7 PERFORMANCE MEASURES INTEGRATING THE SYSTEM OPERATION AND ACTIVITY SYSTEM

The development of performance measures that integrate the activity system relevant to a transit service corridor is one of the main objectives of this project. The performance assessment of a transit system has been historically performed by the agencies that offer the transportation service. Therefore, performance measures have been concentrated on the characteristic of the system, the operation, and in some cases, the interaction with users through user satisfaction surveys or similar instruments. However, one of the main characteristics of the travel demand of any transportation system is that the travel demand is derived from the activities that the system user needs to do. The relation has been clear for many years; however, it has not been widely incorporated in the assessment of the performance of transit systems. The procedure presented in this chapter offers an alternative to consider the integration of the transportation system and the activity system.

This chapter presents various steps that would allow to assess the relationship between the system operation and the relevant activity system. Initially, the stops and passenger loading and unloading are presented. The interaction between a transit system and its users is through the stops along the route. Therefore, a passenger load study was conducted to quantify that relationship along AMA Route 5. The data gathered from the activity system as explained in chapter 4 is then summarized. A statistical correlation analysis helps to determine the variables that are correlated and supports the selection of the variables that are included in the model developed. This model is presented at the end of the chapter.

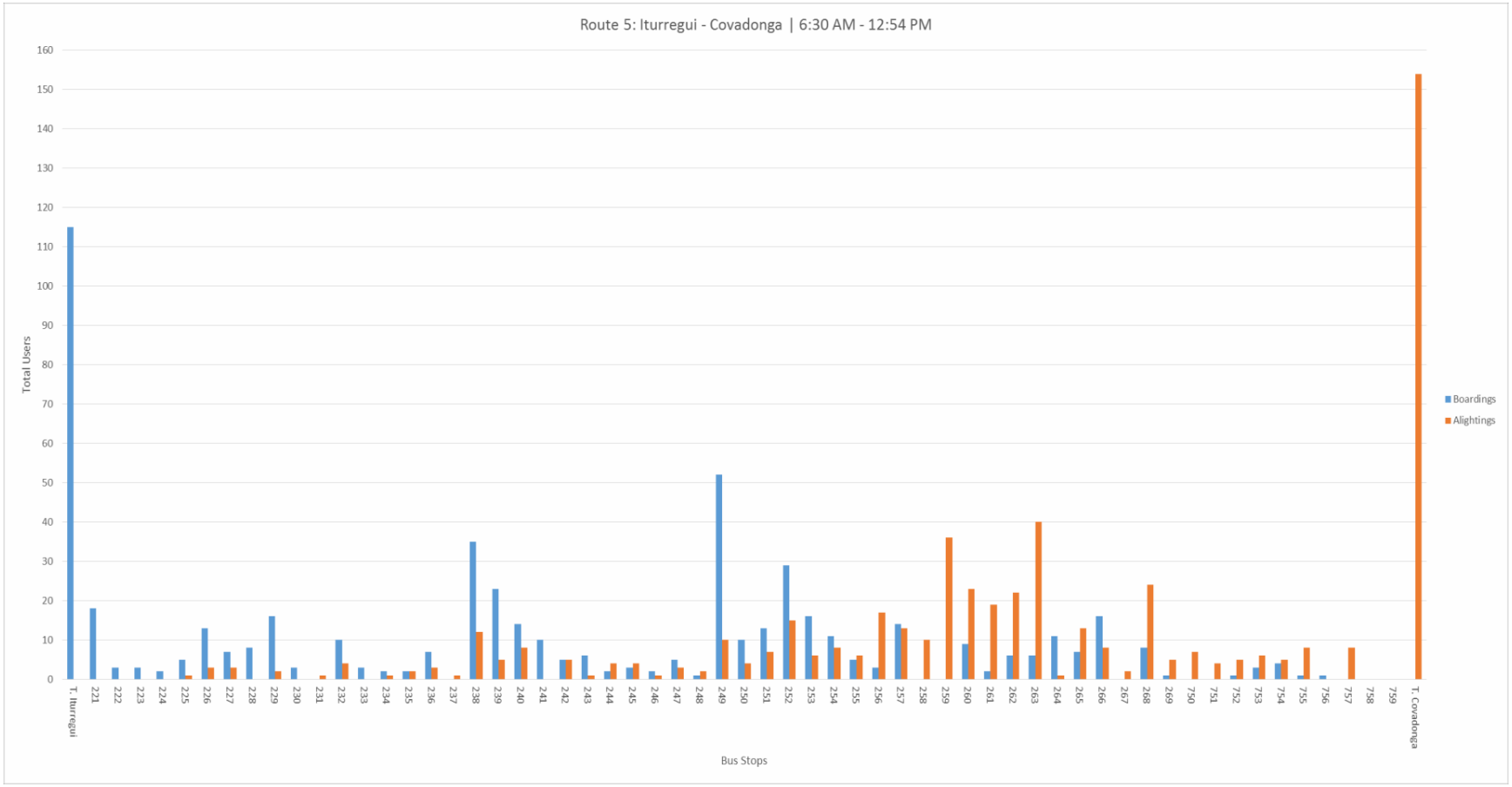
7.1 PASSENGER LOAD STUDY

In order to apply both methodologies to perform all the analyses, it was necessary to perform a boarding and alighting analysis to the AMA Route 5. This boarding and alighting study was conducted on Wednesday March 11, 2015, from 6:30AM to 1:00PM to determine passenger behavior on a typical weekday. The study was limited in scope due to budget and personnel limitations. However, the study allowed us to identify temporal variations and the travel pattern for the peak volume in the morning which is larger than the afternoon's (Vuchic, 2005) mentioned in (Cordero, 2015). Graphic 1 and Graphic 2 show the passenger activities from the Covadonga to Iturregui Terminals and vice versa. For each one of the bus stops, it is shown how many users entered and exited the bus. On Route 5, that day from 6:30am to 1:00pm Table 16 and Table 17, the research team observed a maximum of 338 users between two stops from Iturregui to Covadonga and a maximum of 198 from Covadonga to Iturregui. This means that more than 500 users rode Route 5.



Graphic 1 Graphic of Boarding and Alighting from Covadonga to Iturregui Station

Source: Cruz, W. R. C., Bus Stop Consolidation Analysis for Puerto Rico's Metropolitan Bus Authority, 2015.



Graphic 2 Graphic of Boarding and Alighting from Iturregui to Covadonga Station

Source: Cruz, W. R. C., Bus Stop Consolidation Analysis for Puerto Rico's Metropolitan Bus Authority, 2015.

Table 16 Total Boarding, Alighting and Load by Each Bus Stop from Covadonga to Iturregui Terminal from 6:30AM to 1:00PM

COV-ITU | 6:30 AM - 11:53 AM			
Bus Stop	**Total Boarding**	**Total Alighting**	**Total Load**
T. Covadonga	55	0	55
279	9	0	64
280	2	0	66
281	4	0	70
282	4	0	74
283	2	0	76
284	1	0	77
285	1	0	78
286	0	0	78
287	2	0	80
288	16	0	96
289	2	0	98
290	13	4	107
291	11	1	117
292	8	1	124
44	24	3	145
45	8	3	150
46	15	6	159
47	11	11	159
293	41	5	195
294	8	5	198
295	4	9	193
296	8	8	193
297	5	3	195
298	3	6	192
299	5	9	188
300	12	9	191
301	6	4	193
302	15	13	195
303	4	17	182
304	14	23	173
305	4	10	167
306	1	2	166
307	0	2	164
308	4	7	161
309	1	1	161
310	3	1	163

COV-ITU | 6:30 AM - 11:53 AM			
Bus Stop	**Total Boarding**	**Total Alighting**	**Total Load**
311	1	3	161
312	3	5	159
313	0	4	155
314	0	22	133
315	0	14	119
316	6	25	100
317	0	1	99
318	1	7	93
319	1	0	94
320	4	2	96
321	2	6	92
322	10	6	96
323	1	1	96
324	1	4	93
325	4	1	96
326	0	1	95
327	0	0	95
328	0	2	93
329	0	4	89
330	0	0	89
331	0	11	78
T. Iturregui	0	78	0

Source: Cruz, W. R. C., Bus Stop Consolidation Analysis for Puerto Rico's Metropolitan Bus Authority, 2015.

Table 17 Total Boarding, Alighting and Load by Each Bus Stop from Iturregui to Covadonga Terminal from 6:30AM to 1:00PM

ITU-COV | 6:30 AM - 11:53 AM			
Bus Stop	Total Boarding	Total Alighting	Total Load
T. Iturregui	115	0	115
221	18	0	133
222	3	0	136
223	3	0	139
224	2	0	141
225	5	1	145
226	13	3	155
227	7	3	159
228	8	0	167
229	16	2	181
230	3	0	184
231	0	1	183
232	10	4	189
233	3	0	192
234	2	1	193
235	2	2	193
236	7	3	197
237	0	1	196
238	35	12	219
239	23	5	237
240	14	8	243
241	10	0	253
242	5	5	253
243	6	1	258
244	2	4	256
245	3	4	255
246	2	1	256
247	5	3	258
248	1	2	257
249	52	10	299
250	10	4	305
251	13	7	311
252	29	15	325

ITU-COV | 6:30 AM - 11:53 AM			
Bus Stop	Total Boarding	Total Alighting	Total Load
253	16	6	335
254	11	8	338
255	5	6	337
256	3	17	323
257	14	13	324
258	0	10	314
259	0	36	278
260	9	23	264
261	2	19	247
262	6	22	231
263	6	40	197
264	11	1	207
265	7	13	201
266	16	8	209
267	0	2	207
268	8	24	191
269	1	5	187
750	0	7	180
751	0	4	176
752	1	5	172
753	3	6	169
754	4	5	168
755	1	8	161
756	1	0	162
757	0	8	154
758	0	0	154
759	0	0	154
T. Covadonga	0	154	0

Source: Cruz, W. R. C., Bus Stop Consolidation Analysis for Puerto Rico's Metropolitan Bus Authority, 2015.

7.2 INFLUENCES AREAS AROUND BUS STOPS

As indicated in chapter 4, influence areas around bus stops were drawn and high interest activity places were located with respect to each stop. Considering the Geographical information available from the Census zoning, a map of influence areas around AMA's Route 5 was obtained, as shown in Figure 46.

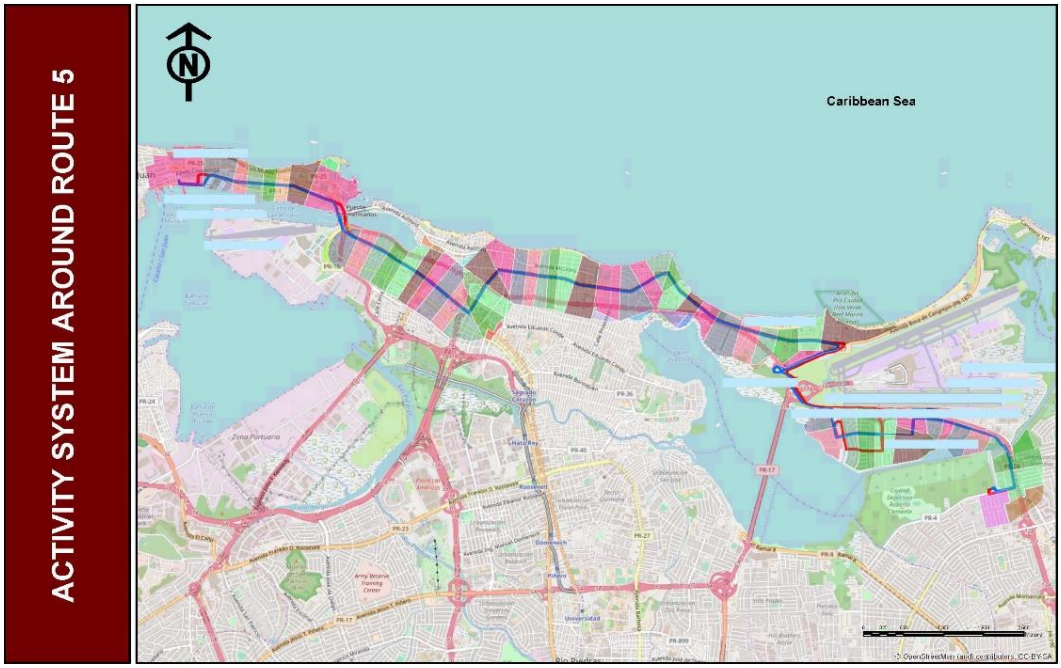


Figure 46 Influence Areas for Activity Systems around Route 5

The influence areas were placed in an Excel table along with the places of interest (i.e., presence or absence of hospitals, schools, governmental offices, industrial zones, recreational zones and touristic zones) and then the correlation analysis was performed as presented below.

Table 18 Consolidate Data for Statistical Analysis

Stop Number	Boarding	Headway Score (HS)	Headway Peak Factor (HPF)	Regularity Index (R)	Inflated Headway Score (IHS)	Income	Distance	Hospital	School	Govern...	Industrial	Recreation	Tourism
3004	115	-1.596	0.358	0.693	-2.303	14,350	595.457	0	0	0	0	0	0
221	18	1.251	0.42	0.687	1.822	14,282.23	426.476	0	1	0	0	0	0
222	3	1.523	0.435	0.71	2.144	18,029.948	337.962	0	1	0	0	0	0
223	3	1.26	0.498	0.728	1.732	16,516.971	362.102	0	0	0	0	0	0
224	2	1.262	0.5	0.728	1.734	13,839	418.429	0	0	0	0	0	0
225	5	1.258	0.502	0.727	1.73	13,835.147	587.411	0	0	0	0	0	0
226	13	1.255	0.503	0.726	1.73	12,666.478	450.616	0	0	0	0	0	0
227	7	1.254	0.502	0.725	1.73	12,455.594	233.355	0	0	0	0	0	0
228	8	1.25	0.499	0.723	1.729	11,037.177	233.355	0	0	0	0	0	0
229	16	1.564	0.483	0.718	2.178	13,447.998	217.261	0	0	0	0	0	0
230	3	1.248	0.491	0.721	1.731	13,781.53	257.495	0	1	0	0	0	0
231	0	1.247	0.491	0.722	1.728	13,682.781	305.775	0	0	0	0	0	0
232	10	1.249	0.484	0.719	1.738	25,221	313.822	0	0	0	0	0	0
233	3	2.622	0.452	0.687	3.818	14,423	362.102	0	0	0	0	0	0
234	2	1.563	0.504	0.749	2.086	13,369.638	265.542	0	0	0	0	0	0
235	2	1.563	0.501	0.748	2.089	13,332.982	217.261	0	0	0	0	0	0
236	7	1.955	0.466	0.733	2.667	12,598.485	1,448.41	0	0	0	0	0	0
237	0	1.306	0.485	0.729	1.79	36,862.075	1,713.951	0	0	0	0	0	1
238	35	1.332	0.476	0.731	1.822	31,191.663	611.551	0	0	0	0	1	1
239	23	1.343	0.479	0.73	1.84	29,797.524	329.916	0	0	0	0	1	1
240	14	1.35	0.482	0.732	1.844	31,798.913	305.775	0	0	0	0	0	1
241	10	1.687	0.405	0.713	2.367	30,019.577	378.196	0	0	0	1	0	0
242	5	1.384	0.484	0.74	1.87	32,600.788	370.149	0	0	0	0	0	1
243	6	1.396	0.487	0.741	1.884	32,383.578	337.962	0	0	0	0	0	1
244	2	1.402	0.489	0.742	1.89	32,482.472	402.336	0	0	0	0	0	0
245	3	1.537	0.497	0.742	2.071	23,137.833	329.916	0	1	0	0	0	0
246	2	1.537	0.496	0.741	2.074	24,614.646	241.402	0	0	1	0	0	0
247	5	1.534	0.495	0.738	2.077	35,470.291	273.588	0	0	0	0	0	0
248	1	1.533	0.491	0.738	2.076	19,258.002	225.308	0	1	0	0	0	0
249	52	1.529	0.489	0.739	2.069	13,960.241	321.869	1	0	1	0	0	0
250	10	1.526	0.486	0.738	2.067	16,509.054	402.336	1	1	0	0	1	0
251	13	1.528	0.486	0.738	2.07	16,822.004	370.149	1	1	0	0	0	0
252	29	1.521	0.485	0.732	2.078	18,756.693	362.102	1	0	0	0	0	0
253	16	1.525	0.487	0.723	2.109	24,144.233	281.635	1	0	0	0	1	0
254	11	1.515	0.49	0.723	2.094	29,016.062	201.168	1	1	0	0	0	1
255	5	1.518	0.491	0.723	2.1	31,371.667	177.028	0	0	0	0	0	0
256	3	1.48	0.493	0.725	2.043	44,962.155	193.121	1	0	0	0	0	0
257	14	1.477	0.488	0.723	2.042	44,359.184	305.775	1	1	1	0	0	1
258	0	1.471	0.487	0.727	2.024	24,249.297	273.588	1	0	1	0	0	0
259	0	1.468	0.488	0.728	2.016	18,906.286	160.934	1	1	0	0	0	0
260	9	1.467	0.49	0.729	2.012	11,733.495	185.075	1	0	0	0	0	0
261	2	1.464	0.494	0.729	2.009	12,978.401	217.261	1	1	0	0	0	0
262	6	1.459	0.496	0.732	1.993	16,325.448	233.355	1	0	0	0	0	0
263	6	1.454	0.5	0.733	1.985	24,303.009	217.261	1	1	0	0	0	0
264	11	1.458	0.498	0.733	1.989	23,615.646	177.028	1	1	0	0	0	0

7.3 CORRELATION ANALYSIS

A correlation analysis was performed between the passengers boardings per stop, the performance measures from the transportation system, and the influence areas and activity places in each area. Section 7.1 presents the details of the boarding/alighting study used to gather the data for the response variable (boardings) used in this correlation analysis. To be consistent with the available data for the response variable, the data for the explanatory variables used on this correlation analysis corresponds to a Wednesday from 6:30 a.m. to 1:00 p.m.

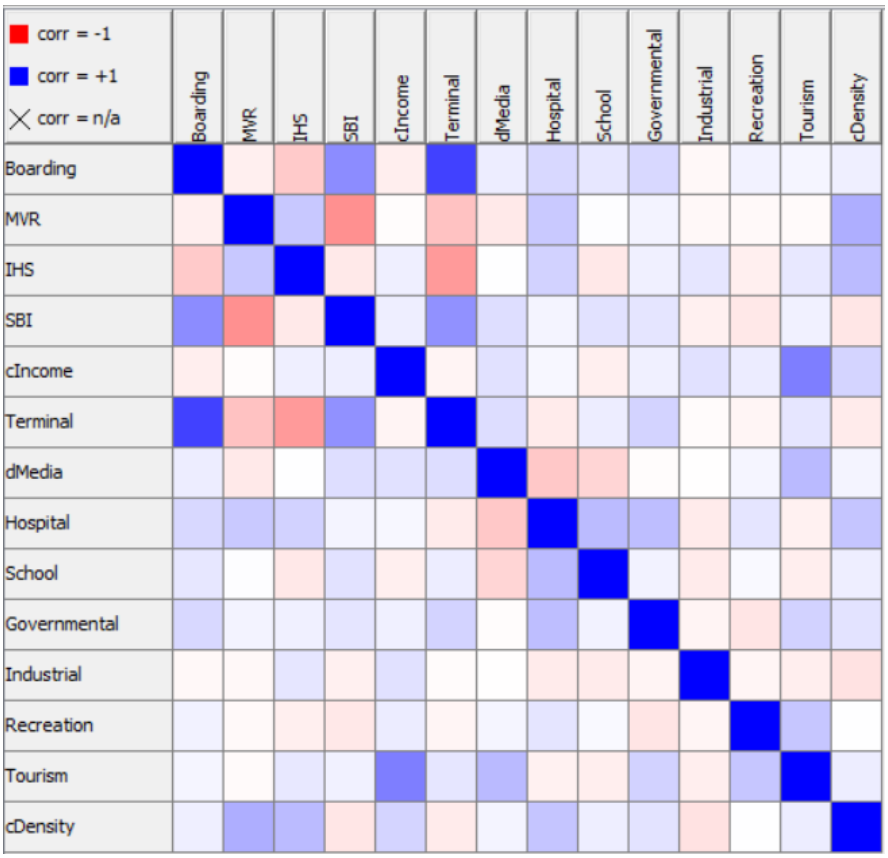


Figure 47 Graphic Correlation Matrix

The correlation analysis between the passengers boarding per stop and the influence areas and activity places in each area revealed a positive linear correlation between the passenger's boarding in a bus stop, and the influence area delimited by the mid-distance to consecutive bus stops and 500 m ahead and behind. Also, there is a negative linear correlation between the passenger's boarding in a bus stop, the mid-income household in the influence area, the regularity measure as the inflated headway score and the mean variation rate. These values indicate that when the headway is too long about the commercial headway, so is the score and the boarding in the bus stop is smaller. The linear correlation between the variables Terminal and Boardings is very high as well. This correlation was expected because in Route 5, the number of boardings at the terminals is high compared with the boardings at immediate stops.

7.4 ACTIVITIES SYSTEM STATISTICAL MODEL

The development of the statistical model incorporating both the transportation and the activity systems requires knowledge and identification of the variables involved. The response variable is the average passenger number that boards a stop with defined characteristics. This variable is a count variable (integer and positive). This type of variable is modeled using an exponential family distribution.

These models have the following form:

$$\mu = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n} \qquad (4)$$

The proposed model intends to define the relationship between the variables that can be explained by Poisson or Negative Binomial distributions:

$$Boarding\ (\mu) = e^{z} \qquad (5)$$

Where:

$$Boarding\ (u) = \beta_0 + \beta_1 Performance + \beta_2 Income + \beta_3 Density + \beta_4 Distance + \beta_5 Hospitals + \beta_6 School + \beta_7 Governmental + \beta_8 Industrial + \beta_9 Recreation + \beta_{10} Tourism \qquad (6)$$

Where:

β_0 *Model Intercept*: Corresponds to the average boarding when all variables are zero. If any variable is centered on the average, when this variable takes values of zero, this value corresponds to the average boarding for the average of the centered variable.

β_1*Performance:* Measurement of the transportation system's performance. This measure may be the frequency, score, or headway uniformity index.

β_2*Income*: Average income per capita of the population living in the stop's influence area. This area is defined as the average mid distance between two consecutive stops, and 500m in both directions, crosswise to the traffic flow of the buses, as shown in Figure 25.

β_3*Density:* Mean populational density of the area around the bus stop.

β_4*Distance:* Average mid distance between consecutive bus stop.

β_5*Hospital:* Dummy variable that indicates the presence or absence of hospitals in the stop's influence area.

β_6*School:* Dummy variable that indicates the presence or absence of schools in the stop's influence area.

β_7*Governmental:* Dummy variable that indicates the presence or absence of governmental entities in the stop's influence area.

β_8*Industrial:* Dummy variable that indicates the presence or absence of industries in the stop's influence area.

β_9Recreation: Dummy variable that shows the presence or absence of recreation areas in the stop's influence area.

β_{10}Tourism: Dummy variable that shows the presence or absence of touristic areas in the stop's influence area.

The typical statistical model used to fit counting variables is the Poisson model. This adjustment depends on the mean and the variance being statistically equal. When this condition does not exist, it is advisable to use another model that captures this dispersion, such as the negative binomial model.

The parameters of the variables were estimated using the Poisson model; however, the estimation of the dispersion yielded a value of 6.454, indicating the existence of significant differences between the mean and the variance. The estimation of the parameters can be seen on Figure 48. This difference between the mean and the variance, made it necessary to use another adjustment model, such as the negative binomial, as shown in R output Figure 49. The calculation of the data dispersion parameter, based on this model, gave a value of 1.29, showing a better fit and a better explanation of the behavior of the data. Appendix D has other estimates made with data corresponding to various dates in the month.

The estimation was made in order to explain the boarding of passengers at each stop, considering the characteristics of the influence area, the activity system in the influence area, and a measure of performance of the transportation system.

The main goal in relating these variables is to identify the degree of association with boarding, in order to identify the magnitude of their contribution to boarding and finally to ridership. Likewise, it is possible to verify whether their contribution to boarding is positive or negative, and if it is statistically significant or not.

There were three groups of explanatory variables: variables specific to the transportation system (performance measure), variables related to the characteristics of the population (population, income), and indicative variables of the activities in the stop's influence area (hospitals, schools, etc.).

The estimated parameters of the negative binomial model are presented in Figure 49. In this model, the average income is measured in thousands and centered on the average, the average distance to the stop is measured in kilometers, the population density centered on the average and the inflated score is expressed in absolute value.

The output of the model estimation process shows the relation between the explanatory variables and the response variable. The variables, inflated score (*IHS*), Stopped Bus Index (SBI), Terminal and Hospital, have a significance level of at least 0.05. The average mid distance (D*istance*) has a positive sign, indicating, in this case, that greater distance between bus stop, implies more passengers boarding per stop. The variable SBI, also, has a positive sign, indicating that the stop where the buses stop the most, are the ones with the highest level of boarding.

A negative sign on the score indicates that as this value increases, the boarding decreases. This is logical, given that a high value of the inflated score indicates that buses are not adhering to the schedule and are arriving before or after their scheduled arrival time. It also indicates that the

arrival of the buses does not occur uniformly. A negative sign in the measured income indicates that as the population has a lower income, the use of the service is higher.

A positive sign in the presence of public institutions, or other institutions or places that in some way generate or attract trips, indicates a positive contribution to the boardings. From the evaluated institutions, the presence of hospitals was the only one reaching significance level. Its significance corresponds solely to the analyzed data. Thus, the study would need to be extended to other routes.

```
Call:
glm(formula = Boarding ~ MVR + IHS + SBI + Terminal + cIncome +
    cDensity + dMedia + Hospital + School + Governmental + Industrial +
    Recreation + Tourism, family = poisson, data = knime.in)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7729  -2.0429  -0.7566   0.6849   8.0818

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.239e-01  6.337e-01   1.142  0.25330
MVR           1.199e+00  6.222e-01   1.926  0.05406 .
IHS          -1.787e+00  2.393e-01  -7.465 8.32e-14 ***
SBI           1.070e+00  4.298e-01   2.490  0.01279 *
Terminal      3.509e+00  1.376e-01  25.495  < 2e-16 ***
cIncome      -1.376e-02  5.561e-03  -2.475  0.01333 *
cDensity      3.946e-05  1.428e-05   2.764  0.00571 **
dMedia       -9.266e-05  2.007e-04  -0.462  0.64430
Hospital      6.487e-01  9.278e-02   6.992 2.71e-12 ***
School        1.828e-02  7.863e-02   0.232  0.81619
Governmental  1.920e-01  1.056e-01   1.819  0.06894 .
Industrial    2.045e-01  3.296e-01   0.621  0.53484
Recreation    2.296e-01  1.145e-01   2.005  0.04501 *
Tourism       1.702e-01  1.181e-01   1.442  0.14936
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1501.81  on 116  degrees of freedom
Residual deviance:  664.78  on 103  degrees of freedom
AIC: 1032.6

Number of Fisher Scoring iterations: 7

[Dispesion] "6.454"
```

Figure 48 R Output for Poisson Statistical Model on April 10 data

```
Call:
glm.nb(formula = Boarding ~ MVR + IHS + SBI + Terminal + cIncome +
    cDensity + dMedia + Hospital + School + Governmental + Industrial +
    Recreation + Tourism, data = knime.in, init.theta = 1.16319006,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4130  -1.0951  -0.3700   0.4024   2.3243

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.797e+00  1.728e+00  -1.040  0.29851
MVR           3.012e+00  1.738e+00   1.733  0.08313 .
IHS          -1.087e+00  4.998e-01  -2.174  0.02971 *
SBI           2.681e+00  1.150e+00   2.332  0.01971 *
Terminal      3.393e+00  7.209e-01   4.707 2.51e-06 ***
cIncome       3.186e-03  1.458e-02   0.218  0.82705
cDensity      5.415e-05  3.675e-05   1.474  0.14059
dMedia        6.883e-05  4.726e-04   0.146  0.88421
Hospital      7.552e-01  2.515e-01   3.003  0.00267 **
School        3.478e-01  2.245e-01   1.550  0.12124
Governmental -2.371e-02  3.414e-01  -0.069  0.94463
Industrial   -2.722e-02  7.871e-01  -0.035  0.97241
Recreation    3.678e-01  3.518e-01   1.046  0.29578
Tourism      -5.128e-02  3.128e-01  -0.164  0.86979
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.1632) family taken to be 1)

    Null deviance: 217.13  on 116  degrees of freedom
Residual deviance: 132.86  on 103  degrees of freedom
AIC: 684.11

Number of Fisher Scoring iterations: 1
              Theta:  1.163
          Std. Err.:  0.199
 2 x log-likelihood:  -654.109

[Dispersion] "1.29"
```

Figure 49 R Output for Negative Binomial Statistical Model on April 10 data

8 APPLYING PERFORMANCE MEASURES TO IMPROVE TRANSPORTATION SYSTEM

An application of the performance measures considering the interaction of the transportation and activity systems is presented in this chapter. Relevant data of the San Juan Metropolitan Area were introduced in previous chapters of this report. These data and all the activity system data for the SJMA public transit system were considered to develop this chapter. The data considered include population and density (using the Census Data), employment, modal split, daily traffic patterns, typical time to initiate home-based trips, travel demand data, levels of service of the SJMA roadway network, general travel times, and current network structure. Besides considering the performance measures presented in previous chapters, the system was modeled using Emme® and TransCAD®. These powerful software applications have the capability to handle all the information for the whole network of this big metropolitan area; however, the Big Data information had to be summarized first using Knime to be used in this project.

The restructuring of the transit network defines actions to improve geographic coverage and reduce walking time. A system of routes hierarchy according to the levels of demand and using the appropriate type and size of vehicles is suggested to reduce waiting times. Exclusive lanes or a combination of mixed traffic with exclusive lanes in peak periods is recommended to improve and regulate the average running speed, therefore, increasing travel time reliability.

This part of the project was performed with the support of TransInfo and the Puerto Rico Integrated Transportation Authority (ATI, for its acronym in Spanish). The source of the figures presented in this chapter is a report submitted to ATI called “Strategies to Increase the Demand of the Public Transit System in the San Juan Metropolitan Area.” (Pipicano *et al.*, 2016)

8.1 STRATEGIES WORKING IN A GENERAL WAY IN THE METROPOLITAN AREA OF SAN JUAN, PUERTO RICO

The Integrated Transportation Authority of the Puerto Rico Department of Transportation and Public Works (DTPW) in the Commonwealth of Puerto Rico is responsible for the planning, administration, management, operation, and maintenance of the integrated public transportation system. ATI’s goal is "Making public transportation the first choice of mobility in Puerto Rico, integrating island-wide public and private services, promoting economic development and quality of life for all." To achieve this vision, ATI has proposed to transform and modernize transportation, so it is reliable, safe, and efficient with regards to socio-environmental and economic sustainability guidelines.

A well designed, efficient and widely used public transportation system is among the most efficient ways to meet the users' travel needs in modern cities. The promotion of public transportation is part of the policies outlined by authorities to mitigate issues of mobility and to promote urban development.

However, public transportation in the Metropolitan Area of San Juan has lacked proper attention to the daily trips of the population. This system serves mainly captive users, who cannot travel by car, and some choice users, who have both the origin and the destination of their trips near the " Tren Urbano" (Urban Train) stations. Tren Urbano is a heavy rail component of the public transportation system with the potential to work as a backbone element for the whole public transportation system in the San Juan Metropolitan Area (SJMA).

Demand for public transportation has declined in recent years, despite efforts made by the authorities to maintain the supply of the system. This is mainly observed in the bus system, which has suffered a general decrease in demand.

The mobility plans that have been developed for the long-term project consist of an integrated network based on the expansion of the metro, tram lines, and BRT-type corridors. The required resources for these projects are very high, and thus their implementation can be delayed, a situation that does not favor the promotion of public transportation endorsed by the city.

This chapter presents short, medium and long-term strategies to restructure the integrated public transportation network for the Metropolitan Area of San Juan considering the performance of the system. The strategies presented allow operational improvement in a short and medium term using fewer resources than those required for rail systems. The corridors will be able to consolidate and, in the future, to evolve to more massive systems.

8.1.1 LONG-TERM VISION

The long-term proposal for the transport network of the Metropolitan Area of San Juan considers the implementation of light rail lines and BRT corridors.

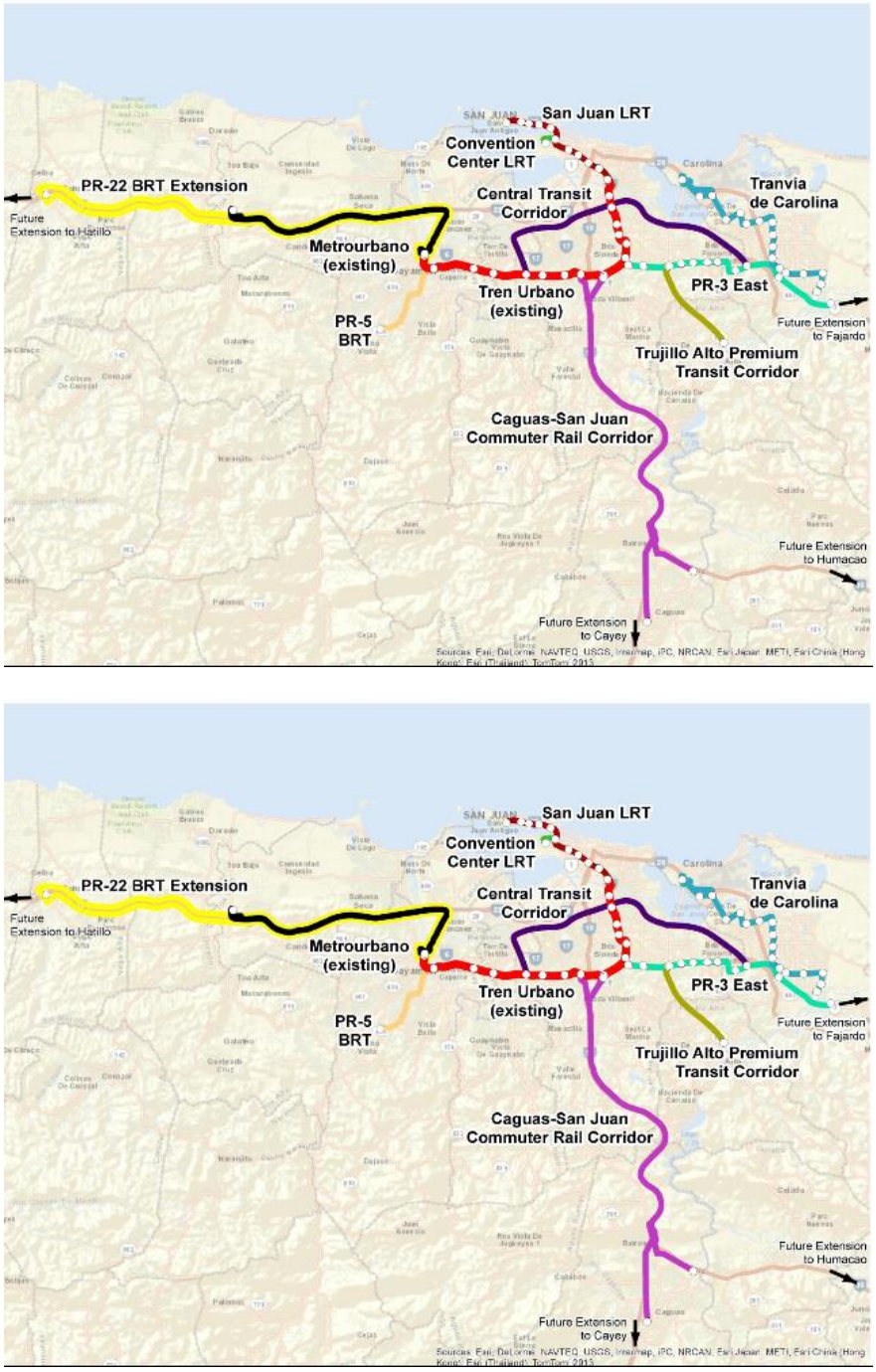


Figure 50 Long-Term Master Plan for Integrated Transport Network

Source: Pipicano, W., I. Ramos, D. Valdés, and C. Figueroa, Estrategias para incrementar la demanda de transporte público colectivo en el Área Metropolitana de San Juan, PR., 2016.

The long-term master plan has forecasted increased demand, which would be reflected in the corridors of the city's transportation network (see Figure 50).

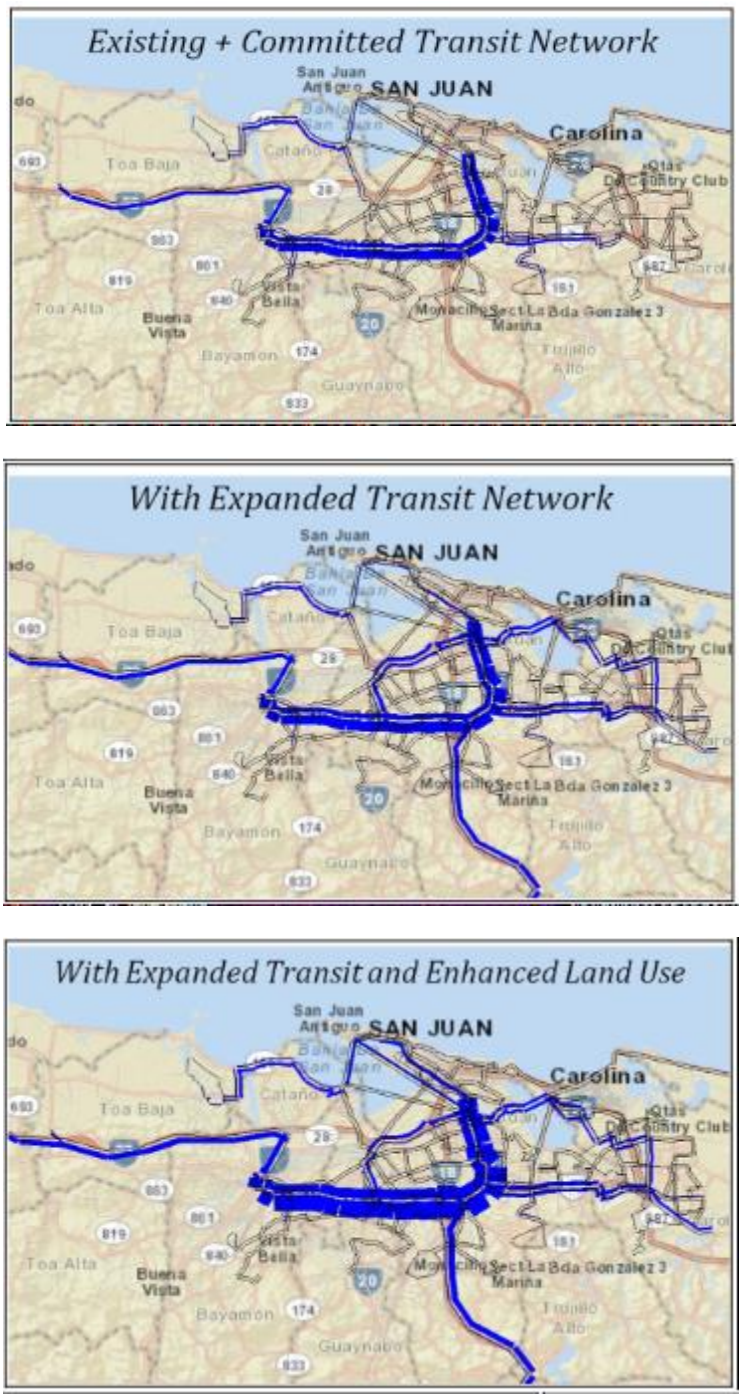


Figure 51 Prognosis of Network Growth and Demand

Source: Puerto Rico Departament of Transportation 2013, as cited in Pipicano, W., I. Ramos, D. Valdés, and C. Figueroa, Estrategias para incrementar la demanda de transporte público colectivo en el Área Metropolitana de San Juan, PR., 2016.

8.1.2 INTEGRATED SHORT-AND MEDIUM-TERM NETWORK

The short- and medium-term proposal consists of structuring a network with two corridors with exclusive lanes, BRT type (Sagrado Corazón Station - Historical Center and Carolina - Rio Piedras) and preferential or exclusive lanes only in peak periods in other corridors of the city. The network includes four transfer terminals (Bayamón, Rio Piedras, Carolina, and Caguas) and four intermediate terminals (Cataño, Martínez Nadal Station, Airport and Art Museum). Figure 52 shows the structure of the proposed network.

Figure 52 Integrated Short-and Medium-Term Network

Source: Pipicano, W., I. Ramos, D. Valdés, and C. Figueroa, Estrategias para incrementar la demanda de transporte público colectivo en el Área Metropolitana de San Juan, PR., 2016.

An essential part of the SJMA transit system is a privately-operated service with fixed routes and small transit vehicles (typically vans or large automobiles) called *Público* that in other cities is called jitneys or "*colectivos.*" Currently, they operate in the suburbs and the city. The short-term strategy recommends that *Público's* routes reach the integration terminals and from there, users would take BRT routes or other integrated bus routes to their destination. The parking and maintenance yards for the fleet operating in the integrated network should be located near the transfer terminals to substantially reduce the downtime and the respective operating costs associated with this part of the operation.

The treatment of preferential corridors for public transport vehicles should be extended to access roads from other municipalities (Caguas, Toa Baja, etc.), as shown in Figure 53.

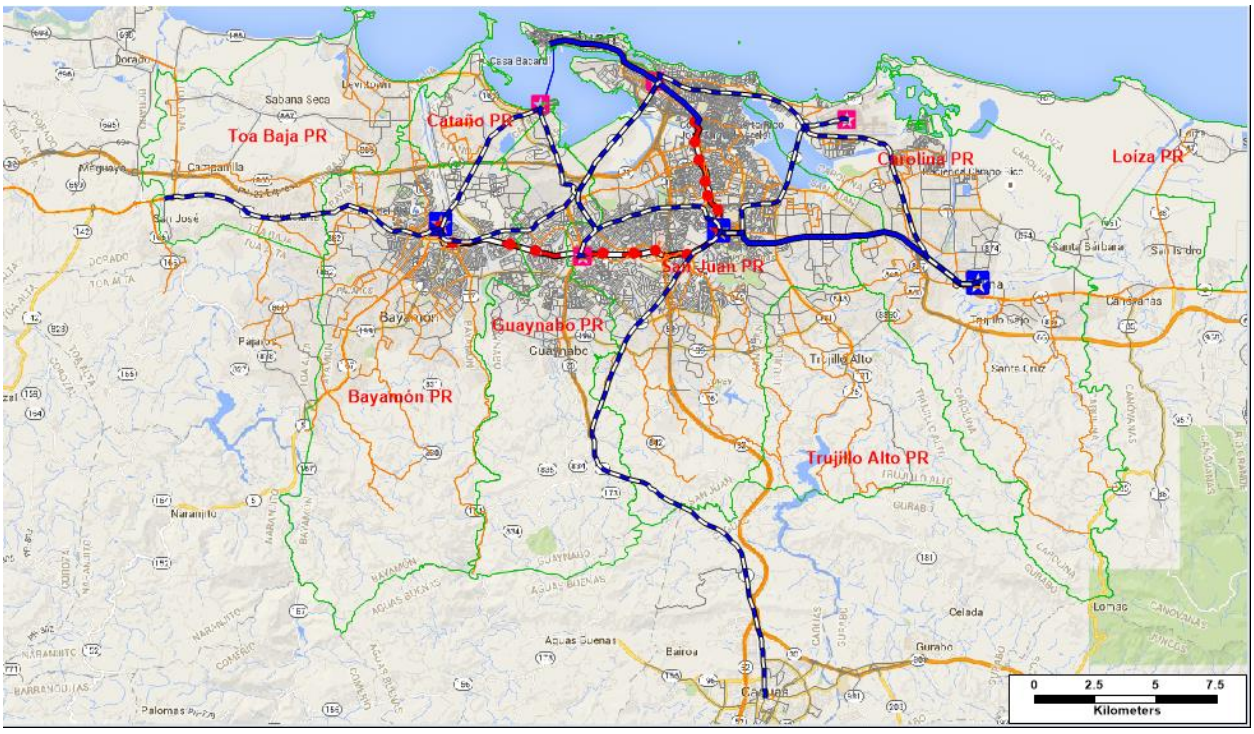


Figure 53 Preferential Corridors for Public Transport Coming from Other Municipalities

Source: Pipicano, W., I. Ramos, D. Valdés, and C. Figueroa, Estrategias para incrementar la demanda de transporte público colectivo en el Área Metropolitana de San Juan, PR., 2016.

Fleet characteristics

For the two BRT corridors, large buses with low entry and doors on both sides were recommended to facilitate the location of the stop stations. In the other routes, vehicles of lower capacity should be considered based on demand and user waiting time.

Attracting passengers to subway stations

The proposed strategy includes adapting bicycle lanes in road corridors to facilitate access to the metro stations. These bicycle riders complement the coverage to areas where there are no integrated network routes. The corridors cover areas up to 2 km from the influence area of the metro stations. In the metro stations, bicycle parking must be created and, better still, parking should be free. Likewise, the alternative of system sponsored public bicycles rented to users with low fares or discount programs should be analyzed.

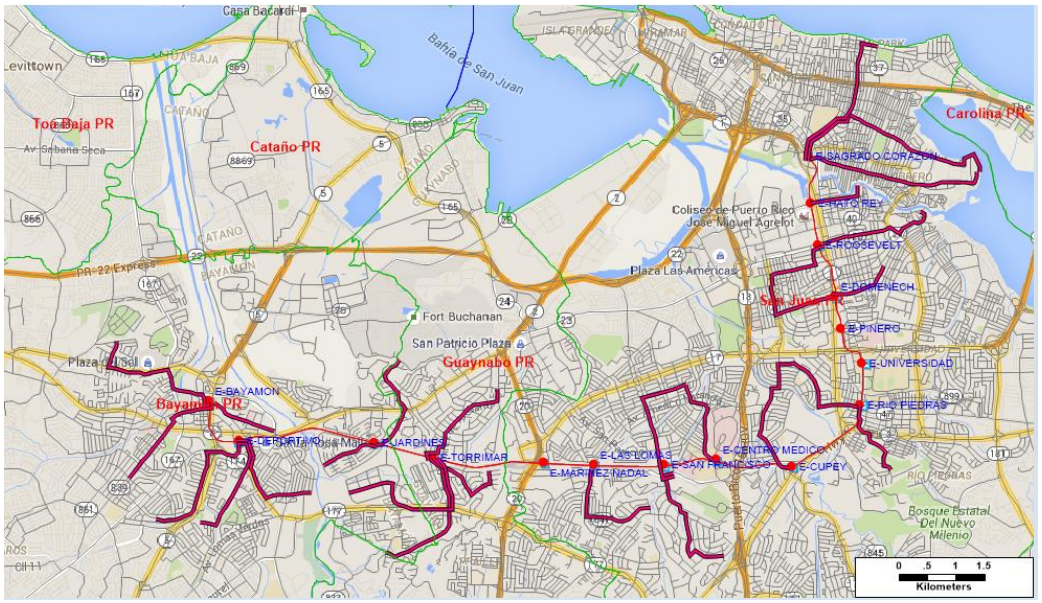


Figure 54 Routes with Cycling Routes to Metro Stations

Source: Pipicano, W., I. Ramos, D. Valdés, and C. Figueroa, Estrategias para incrementar la demanda de transporte público colectivo en el Área Metropolitana de San Juan, PR., 2016.

The strategies developed in this study include four integrated aspects. The basis of the strategy is to give priority to public transportation by having a structured network where the Tren Urbano is the system's structuring axis but requires the extension through BRT express bus systems integrated in trunk corridors and other so-called pre-trunk corridors that initially do not have segregated lanes but are supported by heavier signaling and demarcation to give unity to the system. To complement the integrated network strategy, it is recommended to eliminate bus routes that are parallel to the trunks or pre-trunks and to develop a restructuring of the routes to be covered throughout the Metropolitan Area of San Juan through feeder routes that have a connection with stations of the Tren Urbano or the BRT Systems. Another additional element is the identification of integration terminals −in Bayamón, Rio Piedras, Carolina, and Caguas− in such a way that these integration points allow greater demand consolidation and ease of transfers between feeder and trunk systems. At the same time, it is possible to locate parking and maintenance yards at the ends of the routes, in Bayamón, Carolina and Caguas, to reduce the dead times at the beginning and end of each working day, as well as in periods of transition between Peak and No- Peak periods. Finally, to increase accessibility to the integrated system and the stations of Tren Urbano delineated bicycle paths including cycle paths, shared routes, and bicycle parking in the surroundings of the Train stations are suggested.

The conceptual design proposal of the public transportation system briefly presented in this chapter constitutes an alternative of lower cost and shorter-term implementation than the proposal for the long-term plan. It can serve as a gradual transition and consolidation of the system.

9 CONCLUSIONS AND RECOMMENDATIONS

Transit systems all over the world have been implementing AVL/GPS systems and, in many cases, Automatic Passenger Counters (APC) as well. These data gathering systems generate a massive amount of data per day and therefore offer a great opportunity for using big data methodologies to assess the system performance and improve operations as well as planning.

This project has used off-line AVL/GPS data and activity system's data to develop performance measures integrating not only the transportation system but also the activity system. After overcoming some of the initial obstacles on obtaining the data and deciding what software to use for big data analysis, new performance measures were developed and calculated for one of the AMA routes. Also, transportation and activity systems were considered to assess system performance looking for ways to increase passenger demand for this system. Furthermore, the whole transit network of the San Juan Metropolitan Area was studied, and recommendations drawn for restructuring the network in the short, medium and long-range planning scenarios.

9.1 CONCLUSIONS

This section presents conclusions related to each one of the major sections of this project including (1) development of current performance measures, (2) development of new performance metrics for the transportation system, (3) development of assessment procedures to integrate both the transportation and activity system, and (4) transit network restructuring to improve the current transportation system and increase transit demand.

9.1.1 CURRENT PERFORMANCE MEASURES USING BIG DATA

Two procedures were developed to calculate the typical current performance measures using big data obtained from the AMA AVL/GPS system. The first procedure was developed using MATLAB®; obtaining the results presented in chapter four of this report. Given the complications of using this program, a new process was developed using Knime®, a big data specialized program. This second procedure was used to continue developing the new proposed performance measures using big data.

The following conclusions are presented based on the results obtained by calculating typical performance measures for two of the AMA routes:

- High variations in running time were observed that may be attributed to the fact that buses on the studied routes operate on mixed traffic.
- Routes have a morning peak, an afternoon peak, and a low point at noon.
- Headways were highly variable and ranged in most cases from less than 15 minutes on the lower end (in some extreme cases the headway was less than a minute demonstrating that

bunching is an actual problem) to a maximum of over 80 minutes (with an extremely high value of 394 minutes in one case).

9.1.2 NEW PERFORMANCE MEASURES DEVELOPED

The following conclusions are presented based on the results obtained by calculating the new proposed performance measures for AMA Route 5:

- Minimizing headway variations must be a top priority for AMA.
- The combined effect shown in the new performance measure "inflated headway score (IHS)" is an indication of the degree of dissatisfaction a passenger may experience as a result of uncertainty in waiting for a bus. When a bus does not adhere to the schedule and arrives late but in a uniform fashion, the passenger's dissatisfaction is great but there is certainty that the bus will arrive. However, if in addition to not adhering to the schedule, their arrivals are not uniform and may have very high headways, the dissatisfaction is greater since the bus is not on-time and its arrival time is unknown.
- According to the Inflated Headway Score (IHS), one of the new measures presented in this document, the level of service of AMA Route 5, is between E and F. This is an indicator that buses, on average, do not adhere to the schedule, and arrival times at the stops are not uniform.
- Headway variability may be a consequence of an insufficient fleet size assigned to this route.
- This finding suggests that Headway Adherence must be part of the performance metrics that AMA should monitor on a real-time basis.

9.1.3 ASSESSMENT WITH THE INTEGRATION OF TRANSPORTATION AND ACTIVITY SYSTEMS

The following conclusions are presented based on the results obtained by the correlational analysis and statistical model that have as route boardings a response variable and consider the transportation system's new performance metrics and activity system data as explanatory variables for AMA route 5:

- There is a positive correlation between the passengers boardings in a bus stop, and the area of the influence zone delimited by the mid-distance between consecutive bus stops and 500m around the stop.
- There is a positive correlation between the passengers boardings in a bus stop and the distance of high-interest places located in the influence area.

- The statistical model clearly shows an inversely proportional relationship between the boarding variable and one of the characteristics of the population, specifically the average income.

9.1.4 TRANSIT NETWORK RESTRUCTURING

The following conclusions are presented based on the analysis conducted to propose a complete restructuring of the transit network in the SJMA:

- The long-term network restructuring proposal for the transit network includes the implementation of light rail lines and BRT corridors.
- Short and medium-term strategies include two exclusive BRT corridors (*Sagrado Corazón* Station - Historical Center and *Carolina - Rio Piedras*) and exclusive lanes only in peak periods in other corridors of the city.
- Four transfer terminals (*Bayamón*, *Rio Piedras*, *Carolina*, and *Caguas*) and four intermediate terminals (*Cataño*, *Martínez Nadal* Station, Airport, and Art Museum) are recommended.
- The routes of *Públicos* would be modified to bring the users from the city suburbs to the integration terminals, and from there, users would take other integrated routes of the system (e.g., main routes, secondary routes).
- Maintenance and Parking Yards should be located near the transfer terminals to reduce downtime and associated operating costs substantially.
- Fleet size and type of vehicle evaluations should be conducted before implementing the two BRT corridors. Large buses with low entry doors on both sides should be considered to facilitate the location of the stops and stations. In the other routes, vehicles of lower capacity may be considered depending on the demand and expected performance measures.
- Vehicles should have rails to carry bicycles to facilitate the integration of bike users to the BRT. Bike parking should also be provided at subway stations for integration purposes. The alternative of public bicycles with fare discount programs should also be considered.

9.2 RECOMMENDATIONS

This study was based on data obtained from two routes of the San Juan Metropolitan Bus Authority, and one of them was studied in detail. We recommend expanding the data analysis to the entire system, corresponding to the 30 routes currently in circulation. However, to do that, a new AVL/GPS system is needed. Besides, an APC system would also be of great importance to obtain real-time passenger data in all routes that can generate the information required to improve the knowledge of this transit system. A more detailed study of bus stops is recommended to identify the activity system characteristics on the stops' influence area and determine what kind of combinations produce high boarding numbers in the transit system.

Even though Knime® is designed to work with big data, looking for other big data management software that also incorporate geographical information is recommended to be able to represent the traffic data and performance measures as layers on top of the map of the covered area. This type of application would be beneficial for non-technical decision makers that could experience the performance of the system easier with better graphical representations.

Although the coefficients of the variables related with the activity system were not significant, except for the variable "Hospital," they were left in the model in order to show their contribution to the passengers' travel. It is recommended to expand the sample to increase the number of areas that have travel generators and perform additional analysis to relate travel generators with actual stop boardings.

10 REFERENCES

Transportation Research Board, 2010. *TCRP Report 141: A Methodology for Performance Measurement and Peer Comparison in the Public Transportation Industry,* Whashington DC: Transportation Research Board.

(FHWA), T. F. H. A., 2018. *Travel Time Data Collection Handbook,* s.l.: s.n.

Anon., n.d. s.l.:s.n.

Auckland Transport, 2014. *Bus Reliablity and Puntuality Performance.* s.l., Auckland Transport, p. 6.

Berkow, M., Chee, J., Bertini, R. L. & Christopher, M., 2007. *Transit Performance Measurement and Arterial Travel Time Estimation Using Archived AVL data.* Portland, s.n., p. 10.

Board., Transportation Research, 2009. *TCRP Report 136: Guidebook for Rural Demand-Response Transportation: Measuring, Assessing, and Improving Performance,* Washington D.C.: Board., Transportation Research.

Cevallos, F. & Wang., X., 2008. ADAMS: Data Archiving and Mining System for Transit Service Improvements. *Transportation Research Record: Journal of the Transportation Research Board, No. 2063,,* p. pp. 43–51..

Cordero, W. R., 2015. *Bus Stop Consolidation Analysis for Puerto Rico's Metropolitan Bus Authority,* Puerto Rico: Universidad de Puerto Rico.

Cortés, C. et al., 2011. Comercial bus speed diagnosis based on GPS-Monitored data. *Elsevier,* pp. 695-707.

Crout, D. T., 2011. *Using Integrated Data to Measure Performance at TriMet, Conference.* Oregon, s.n.

Cruz, W. R. C., 2015. *BUS STOP CONSOLIDATION ANALYSIS FOR PUERTO RICO'S METROPOLITAN BUS AUTHORITY,* Puerto Rico: s.n.

Czech, P. & Turner, S., 2014. *Using GPS Data for arterial Mobility Performance Measures.* San Paul, Minesota, Mn DOT y TTI, p. 55.

Demiryurek, U., 2016. Trajectory Data Mining for Performance Measurement of Public Transportation System.

Federal Transit Administration, U.S Department of Transportation., 2015. *National Transit Database. Policy Manual. Office of Budget and Policy. ,* Washington D.C.: U.S Department of Transportation.

Furth, P., Hemily, B., Muller, T. & Strathman, J., 2006. *TCRP Report 113: Using Archived A VL-APC Data to Improve Transit Performance and Management*, Washington, D.C.: Transportation Research Board of the National Academies.

Gokasar, I. & Simseck, K., 2014. *Using "Big Data" For Analysis and Improvement of Public Transportation System in Istanbul.* Stanford, ASE, p. 7.

Henderson, G., Kwong, P. & Adkins, H., 1991. *Regularity Indices for Evaluating Transit Performance*, s.l.: Transportation Research Record 1297.

Hernández, R., 2006. *Metodologia de la Investigación.* Mexico, D.F.: MacGraw Hill.

Hickey, A. M., Nuworsoo, C. & Pangilinan, C. A., 2014. Dynamic Dispatch with Advanced Train Control System Data: Application to Muni Metro Light Rail Vehicles Departing Embarcadero Station in San Francisco, California.. *Transportation Research Record: Journal of the Transportation Research Board, No. 2415* , pp. pp 35-47.

Ji, Y. & Zhang, M., 2013. Dynamic Holding Strategy to Prevent Buses from Bunching. *Transportation Research Record: Journal of the Transportation Research Board, No. 2352,*, pp. pp 94-103.

Kittelson & Associates, I., Transportation, I. T. & Transport Consulting, L., 1999. *TCRP Web-Only Document 6: Transit Capacity and Quality of Service Manual, 1st Edition.*, Washington, D.C: Transportation Research Board, National Research Council.

Liao, C.-F. & Liu, H. X., 2010. Development of Data-Processing Framework for Transit Performance Analysis. *Transportation Research Record: Journal of the Transportation Research Board No. 2143*, p. pp. 34–43. .

Lock, O. & Erhardt, G., 2015. *Keeping Track-The fusion of large, automatically-collected transport data in capturing long-term system change.* Brisbane, Australia, s.n.

Muller, T. & Furth., P. G., 2001. Trip Time Analyzers, Key to Transit Service Quality. In. *Transportation Research Record: Journal of the Transportation Research Board, No. 1760*, p. pp. 10–19.

Munizaga, M., 2015. *Harnessing Big Data, How Big Data can be leveraged to improve urban transportation and quality life.* Santiago de Chile, s.n., p. 57.

Saavedra, M., Hellinga, B. & Casello, J., 2011. Automated Quality Assurance Methodology for Aechived Transit Data from Automatic Vehicle Location and Passenger Counting Systems.. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2256(No. 2256), p. pp. 130–141..

Salek, S., Noroozi, R., Casello, J. M. & Hellinga., B., 2011. Predicting the Mean and Variance of Transit Segment and Route Travel Times.. *Transportation Research Record: Journal of the Transportation Research Board, No. 2217,* p. pp. 30–37.

Segarra Algueró, P., 2013. *Using Smart Card Thechnologies to Measure Public Transport Performance: Data Capture and Analysis,* Spain: Universitat Politècnica de València.

Shannon, E. & Bellisio, A., 2013. *The MTA in the Age of Big Data: Transforming the Wealth of MTA Data into Accessible, Meaningful, Visual, Interactive Information. ,* New York: Metropolitan Transportation Authority.

Shi, J., Sun, Y., Schonfeld, P. & Guo, Q., 2015. *Identifying Passenger Flow Characteristics and Evaluating Travel Time Reliable by Visualizing AFC Data: A Case of Study of Shanghai Metro.* Washington D.C. , s.n.

Steward, C., Diab, E., Bertini, R. & El-Geneidy, A., 2015. *Perspectives on transit: Potencial benefits of visualizing transit data.* Cambridge, MA USA, s.n.

Transportation Research Board., 2003. *TCRP Report 88: A Guidebook for Developing a Transit Performance-Measurement System,* Washington : Transportation Research Board..

Transportation Research Board., 2013. *TCRP Report 165: Transit Capacity and Quality of Service Manual. Third Edition,* Whashington D.C.: Transportation Research Board..

Transportation Research Board, 1998. *Transit Scheduling: Basic and Advanced Manuals TCRP Report 30,* Washington, D.C.: NATIONAL ACADEMY PRESS.

Transportation Research Board, 2003. *TCPR Synthesis 56: Performance-Based Measures in Transit Fund Allocation,* Washington, D.C.: Transportation Research Board.

Transportation Research Board, 2010. *Higway Capacity Manual,* Washington D.C.: Transportation Research Board.

Transportation Research Board, 2017. *TRB Straight to Recording for All: Using Automated Transit Data to Manage Operations and Improve System Performance.* [Online] Available at: http://www.trb.org/ElectronicSessions/Blurbs/176060.aspx

Tribone, D. et al., 2014. Automated, Data-Driven Performance Regime for Operations Management, Planning, and Control. *Transportation Research Record: Journal of the Tr,* pp. pp 72-79.

Van Oort, N. & Cats, O., 2015. *Improving publics transport decision making, planning and operations by using Big Data: Cases from Sweden and Netherlands.* Washington, DC, USA, s.n., pp. Pages 19-24.

Berkow, M. *et al.* (2007) 'Transit performance measurement and arterial travel time estimation using archived AVL data', *ITE District,* 6.

Carter, D. N. and Lomax, T. J. (1992) 'Development and application of performance measures for rural public transportation operators', *Transportation Research Record*, 1338.

Chaves, G. and Hernández, H. (2015) 'Desempeño y calidad de servicio de autobuses externos de la Universidad de Costa Rica'. Programa Infraestructura del Transporte (PITRA), LanammeUCR.

Chu, X. (2004) 'Ridership Models at the Stop Level. National Center of Transit Research'.

Corporation, I. B. M. (2011) 'IBM SPSS Modeler CRISP-DM Guide'.

Evans, I. V and others (2004) 'Transit Scheduling and Frequency-Traveler Response to Transportation System Changes', *Transportation Research Board. 2004.*

Faraday, J. (2009) *Linear Model with R*. Boca Raton London New York: Chapman & Hall/CRC.

Gutiérrez, J., Cardozo, O. D. and García-Palomares, J. C. (2011) 'Transit ridership forecasting at station level: an approach based on distance-decay weighted regression', *Transport Geography*, Vol. 19(6), pp. 1081–1092.

Navarro Díaz, C. (2017) 'La geografía de la desigualdad y el transporte colectivo en Puerto Rico', in *II Conferencia Hermenegildo Ortiz Quiñones sobre 'Movilidad y Equidad: Retos para la Planificación y Política Pública'*. San Juan, Puerto Rico.

Pipicano, W. *et al.* (2016) 'Estrategias para incrementar la demanda de transporte público colectivo en el Área Metropolitana de San Juan, PR.'

Rodríguez Marrero, M. L. (2016) 'Expertos analizan el transporte público y colectivo en Puerto Rico'.

TRB, Transportation Research Board (2012) *Information for Authors: a guide for preparing and submitting manuscripts for presentation at the TRB Annual Meeting and for Publication in TRB's Journal*. Washington, D.C.

Widhalm, P. *et al.* (2015) 'Discovering urban activity patterns in cell phone data', *Transportation*, 42(4), pp. 597–623. doi: 10.1007/s11116-015-9598-x.

11 APPENDIX

11.1 APPENDIX A

11.1.1 BUS STOP INFORMATION GATHERING

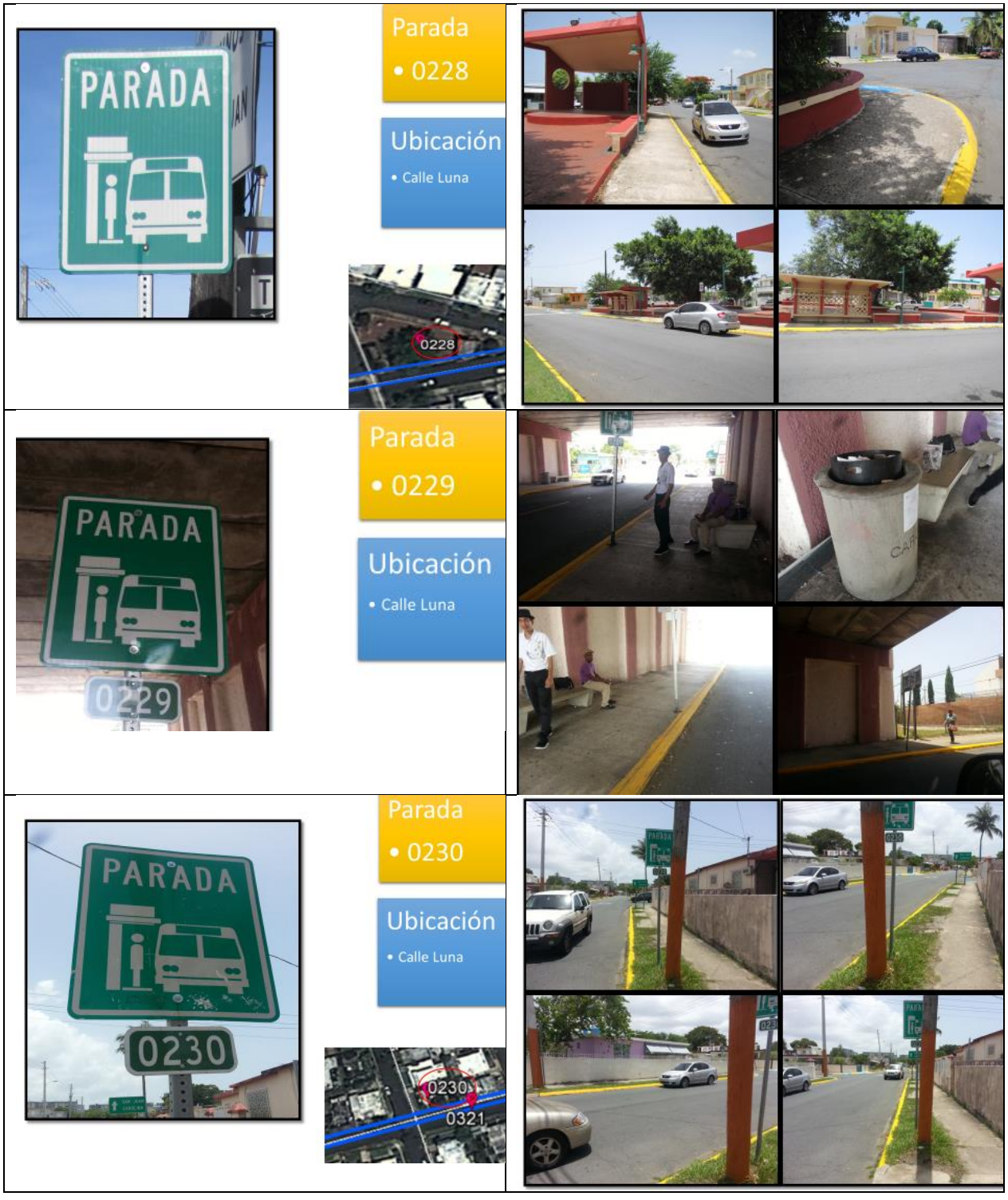


Figure 55 Stops` appraisal sample of Route 5 AMA

Figure 56 Stops` appraisal sample of Route 5 AMA (cont.)

11.2 APPENDIX B

11.2.1 KNIME PROCESS OF DATA MINING AND ANALYSIS

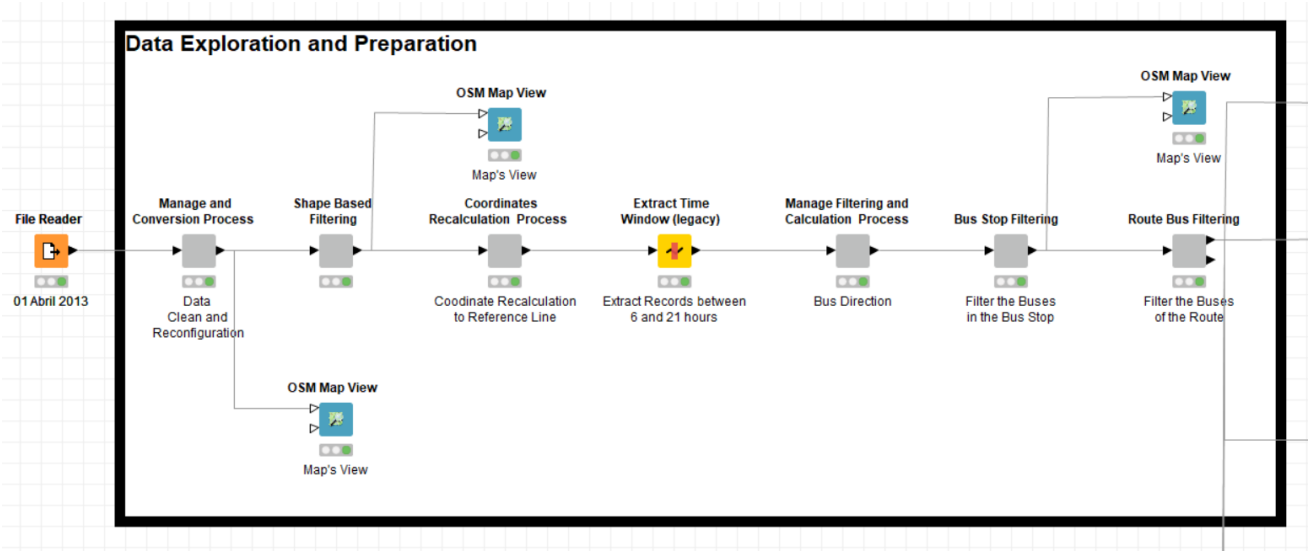


Figure 57 Data Exploration and Preparation Process

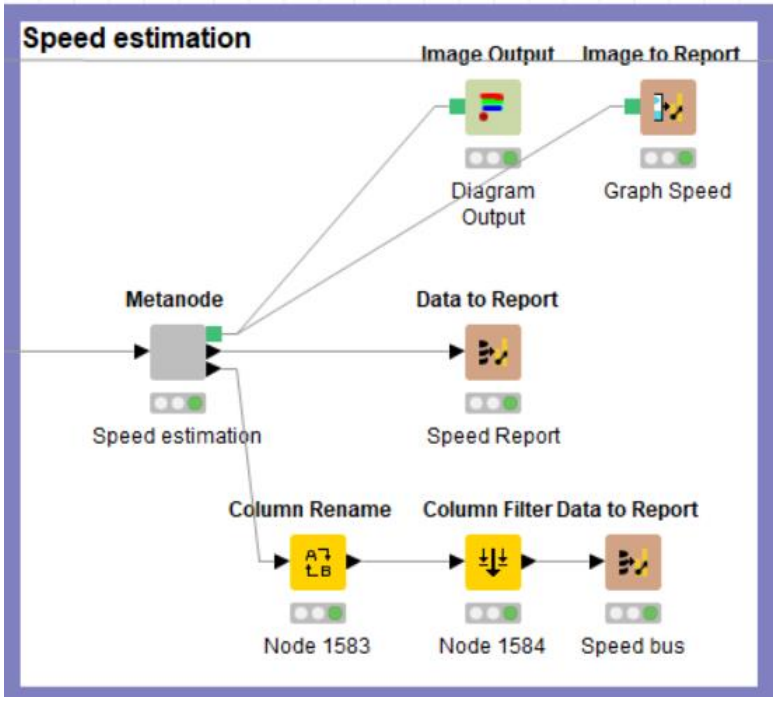


Figure 58 Speed Estimation Process

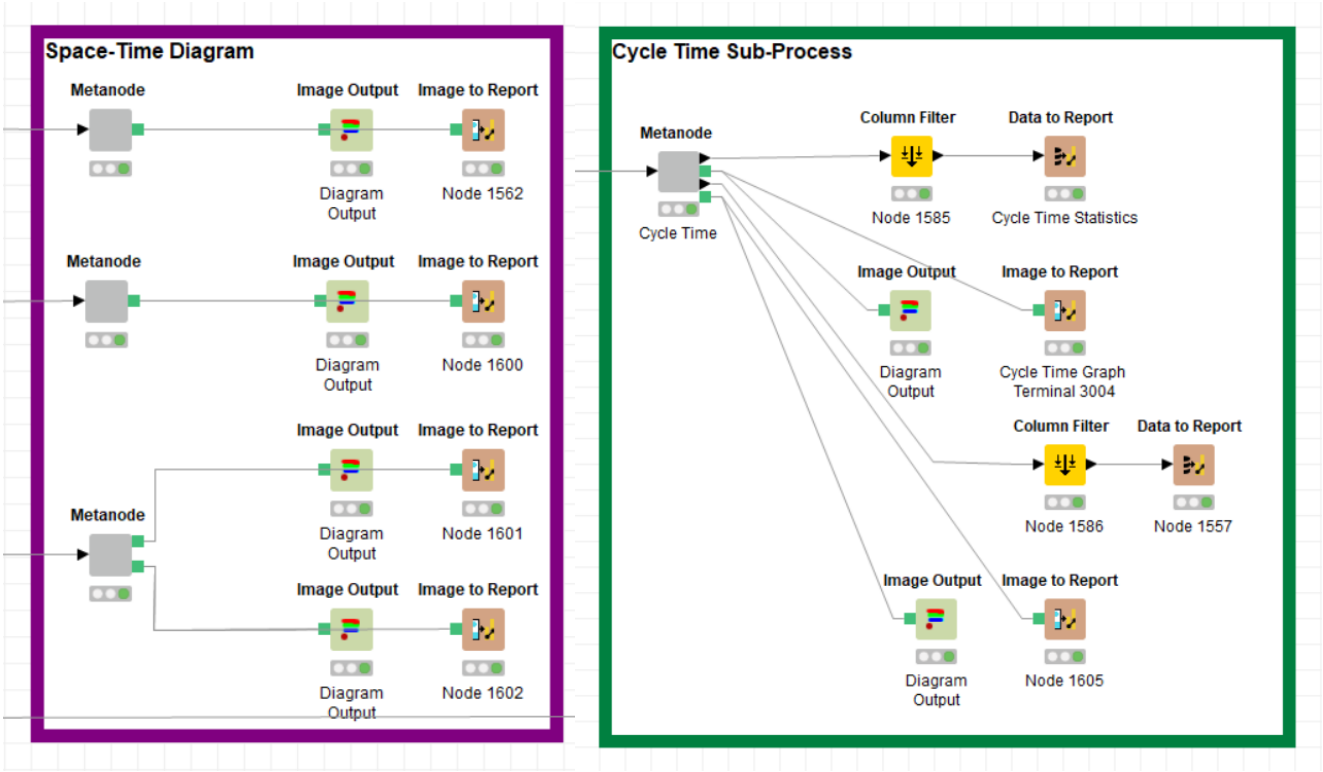


Figure 59 Space-Time and Cycle-Time Process

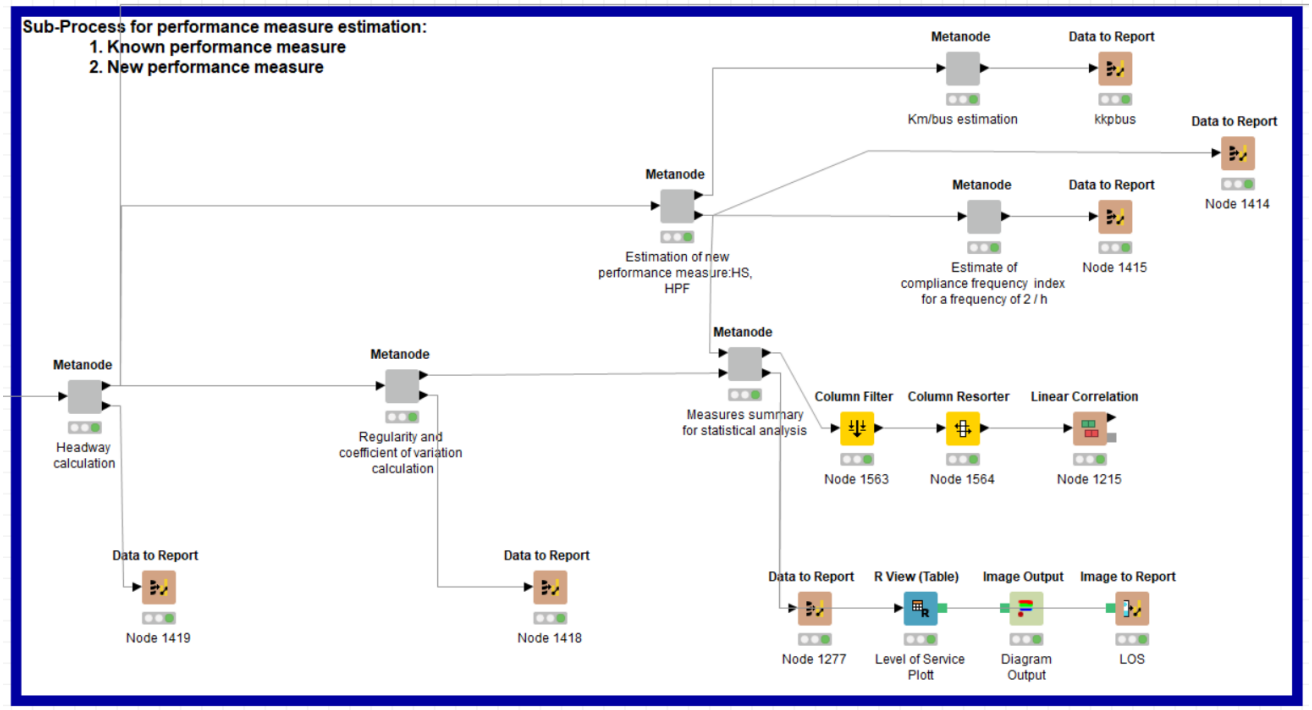


Figure 60 Performance Measure Estimation Process

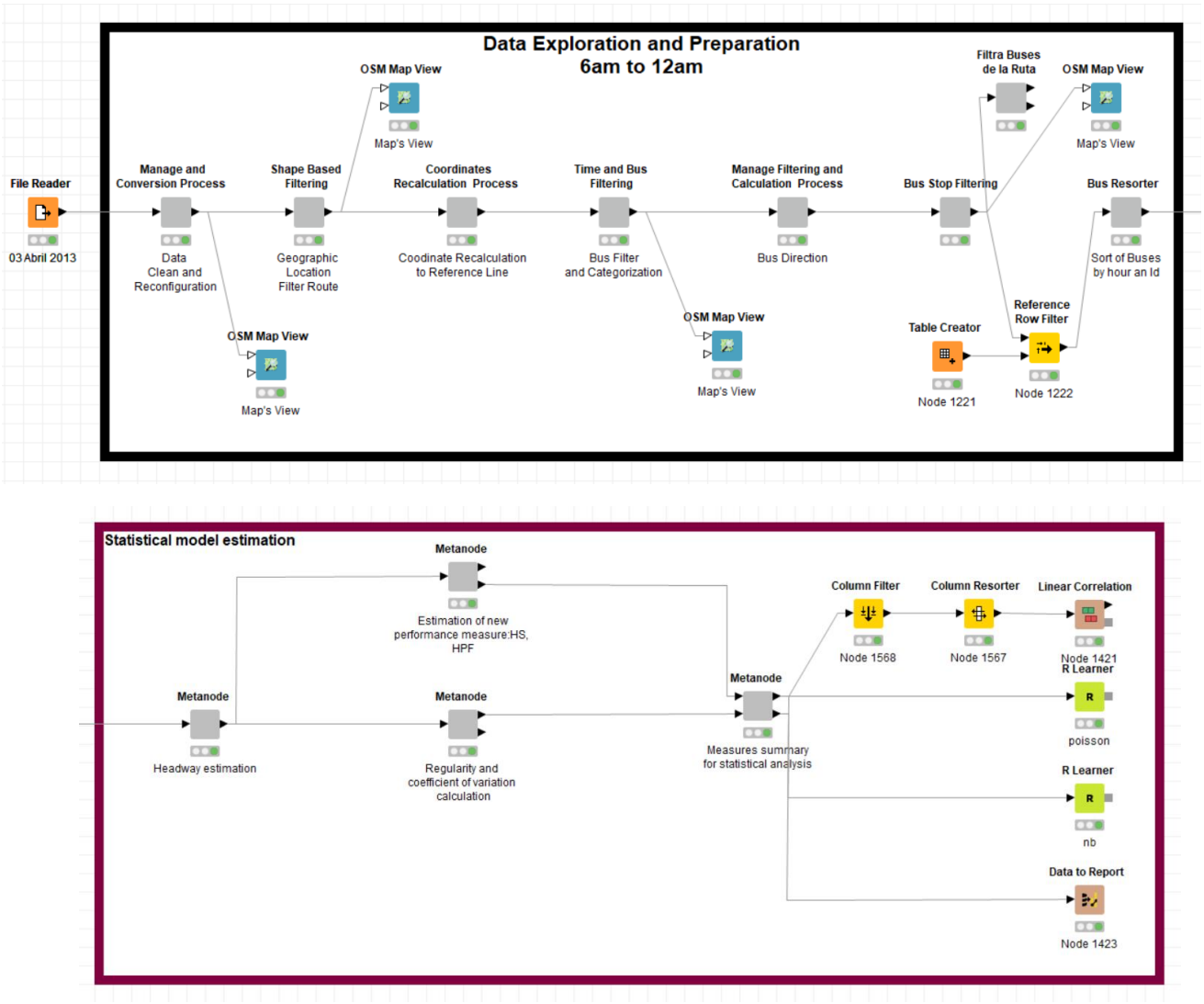
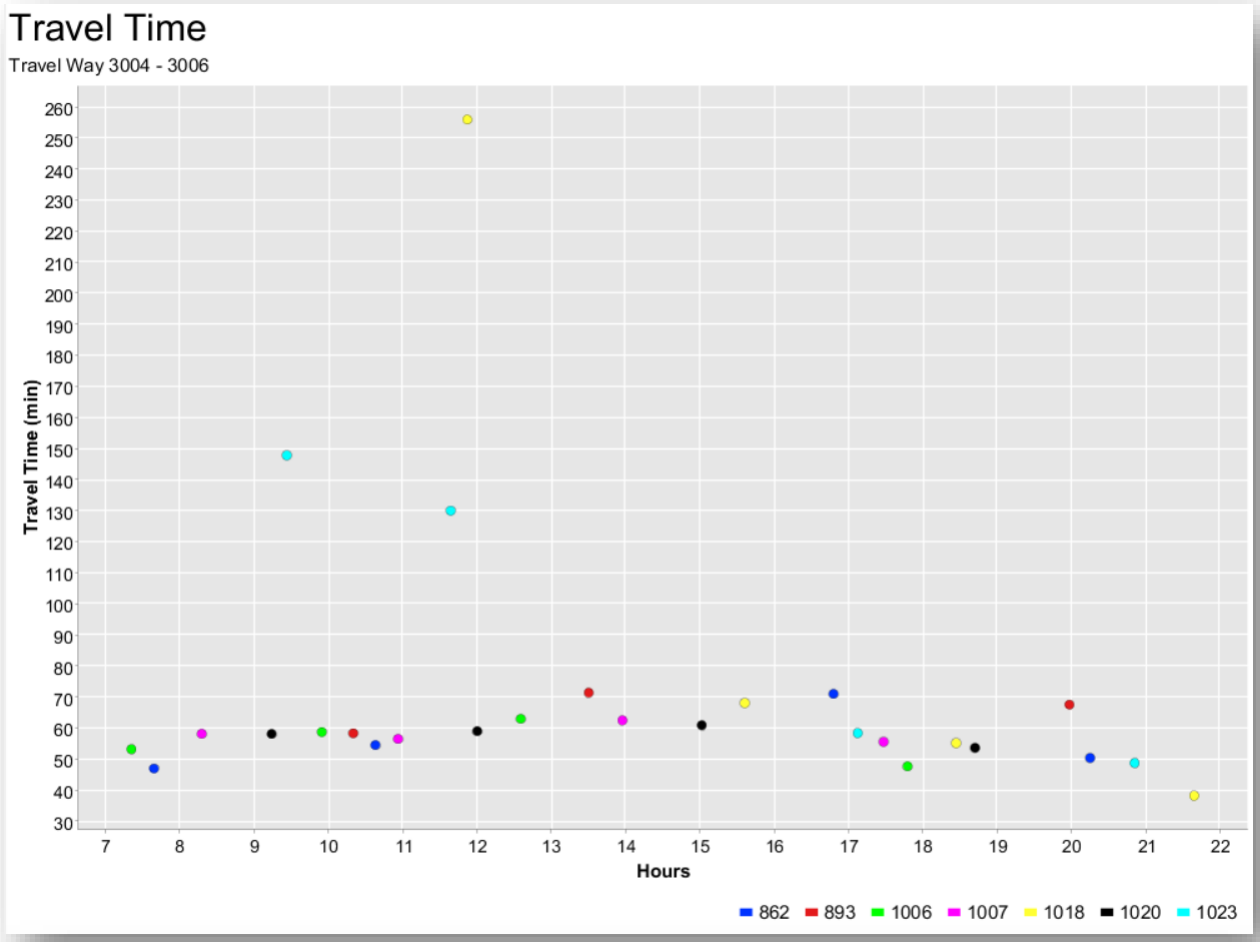


Figure 61 Statistical Model Estimation Process

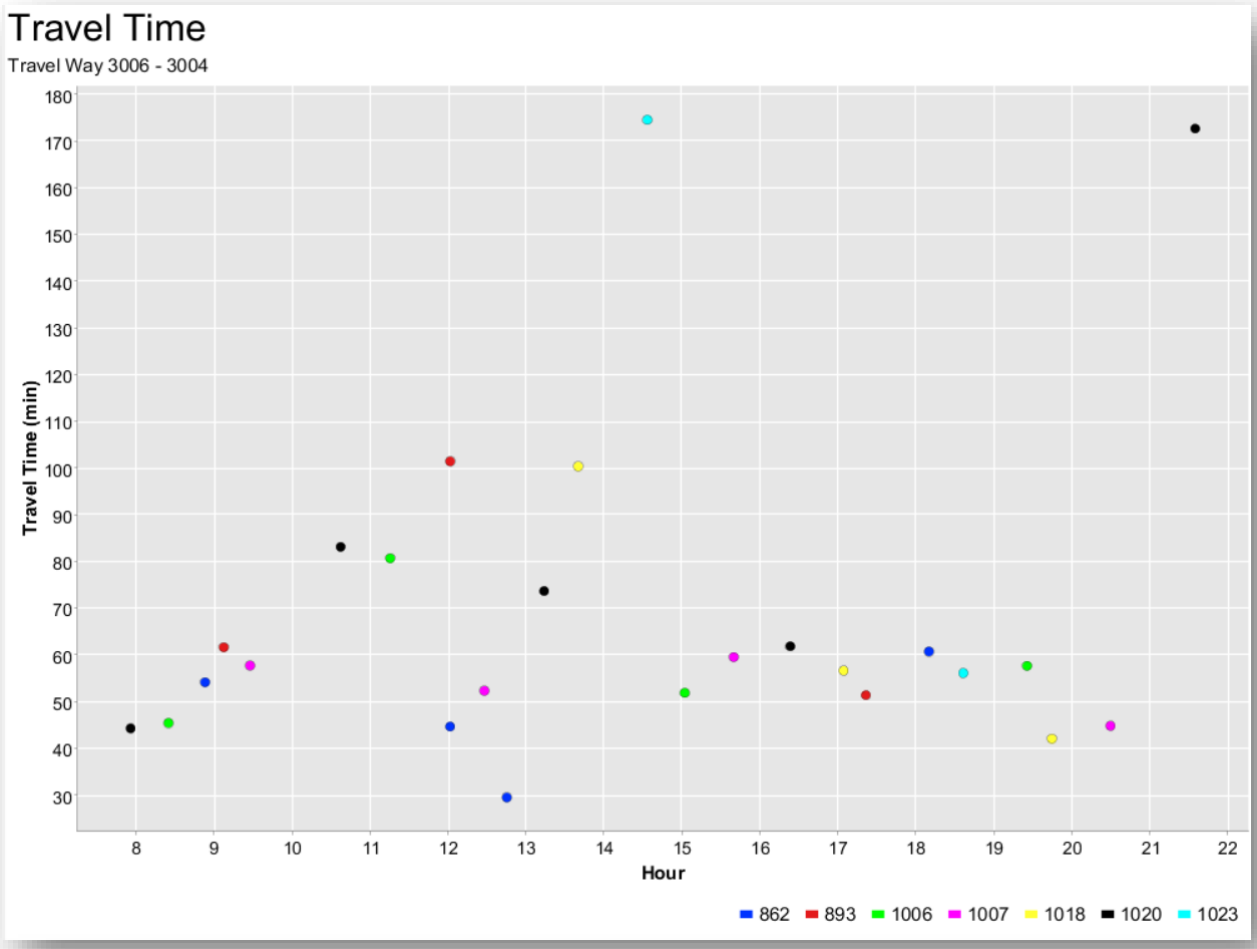
11.3 APPENDIX C

11.3.1 ONE DAY TRAVEL TIME



Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Histogram
TravelTimeMin	38.167	147.75	63.562	23.543	554.26	2.837	8.248	38 148

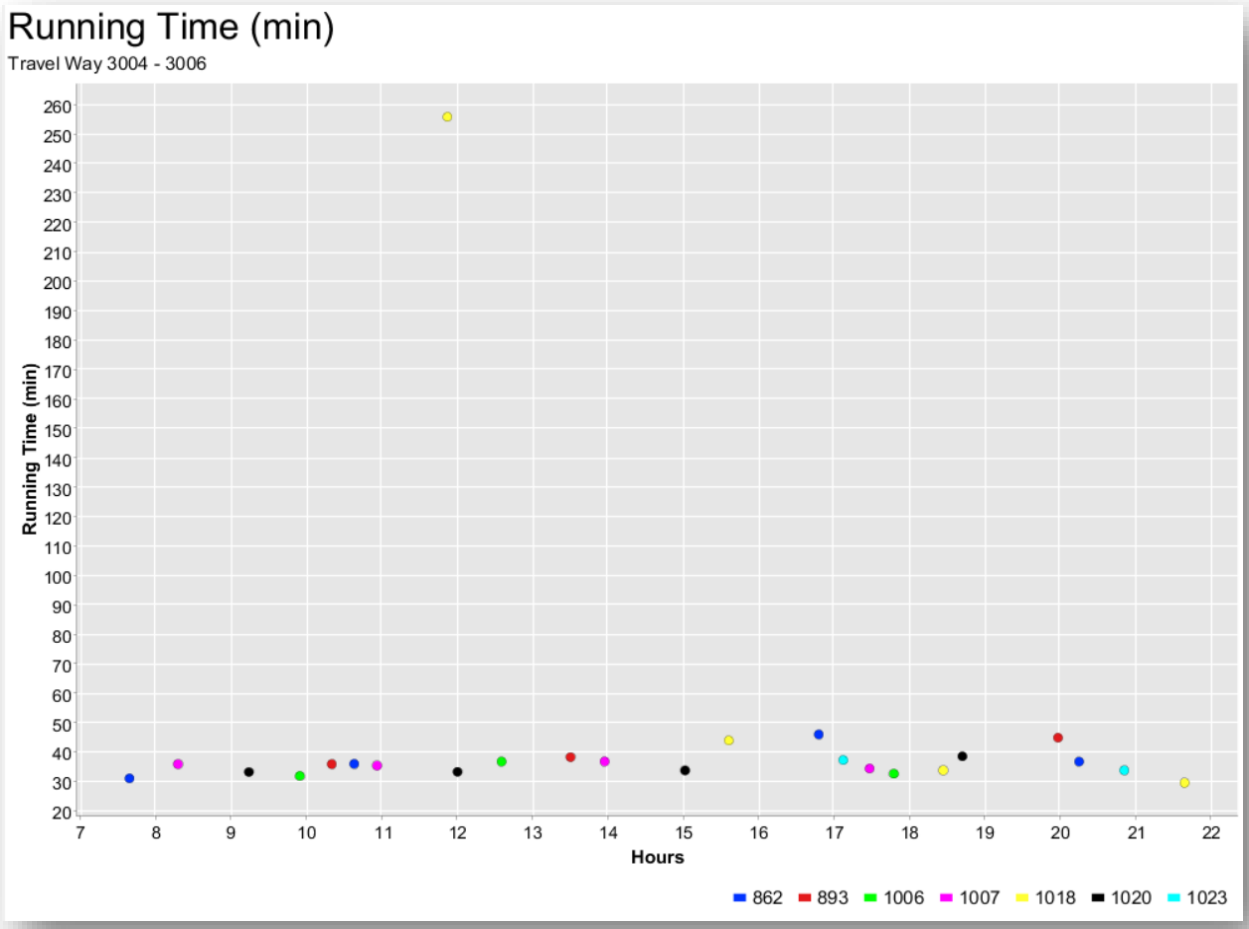
Figure 62 One-way travel time 3004 to 3006



Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Histogram
TravelTimeMin	29.45	101.417	59.602	17.774	315.91	1.047	1.017	29 101

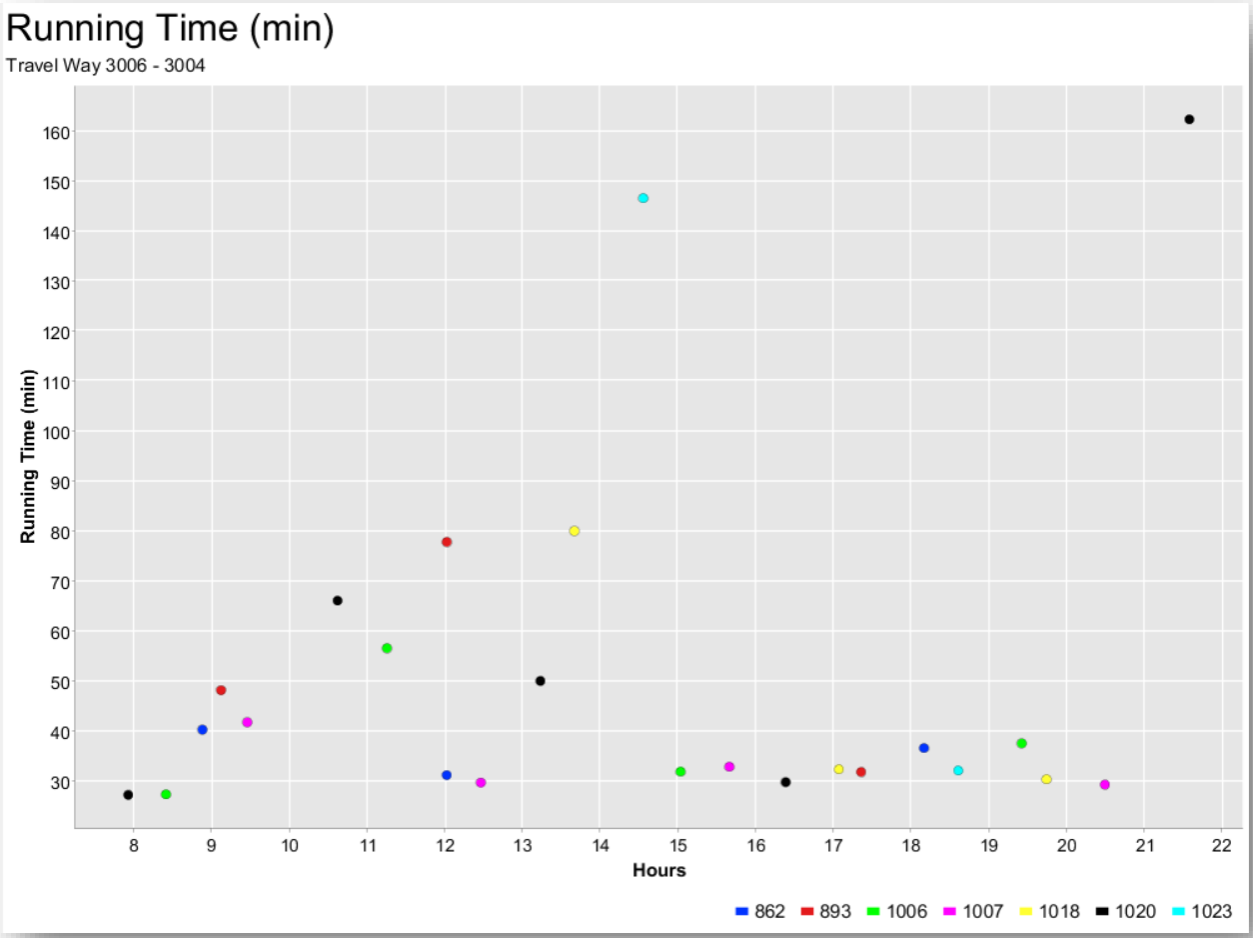
Figure 63 One-way travel time 3006 to 3004

11.3.2 ONE DAY RUNNING TIME



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
RunningTime	29.5	45.967	36.053	4.181	17.484	1.069	0.989	29 46

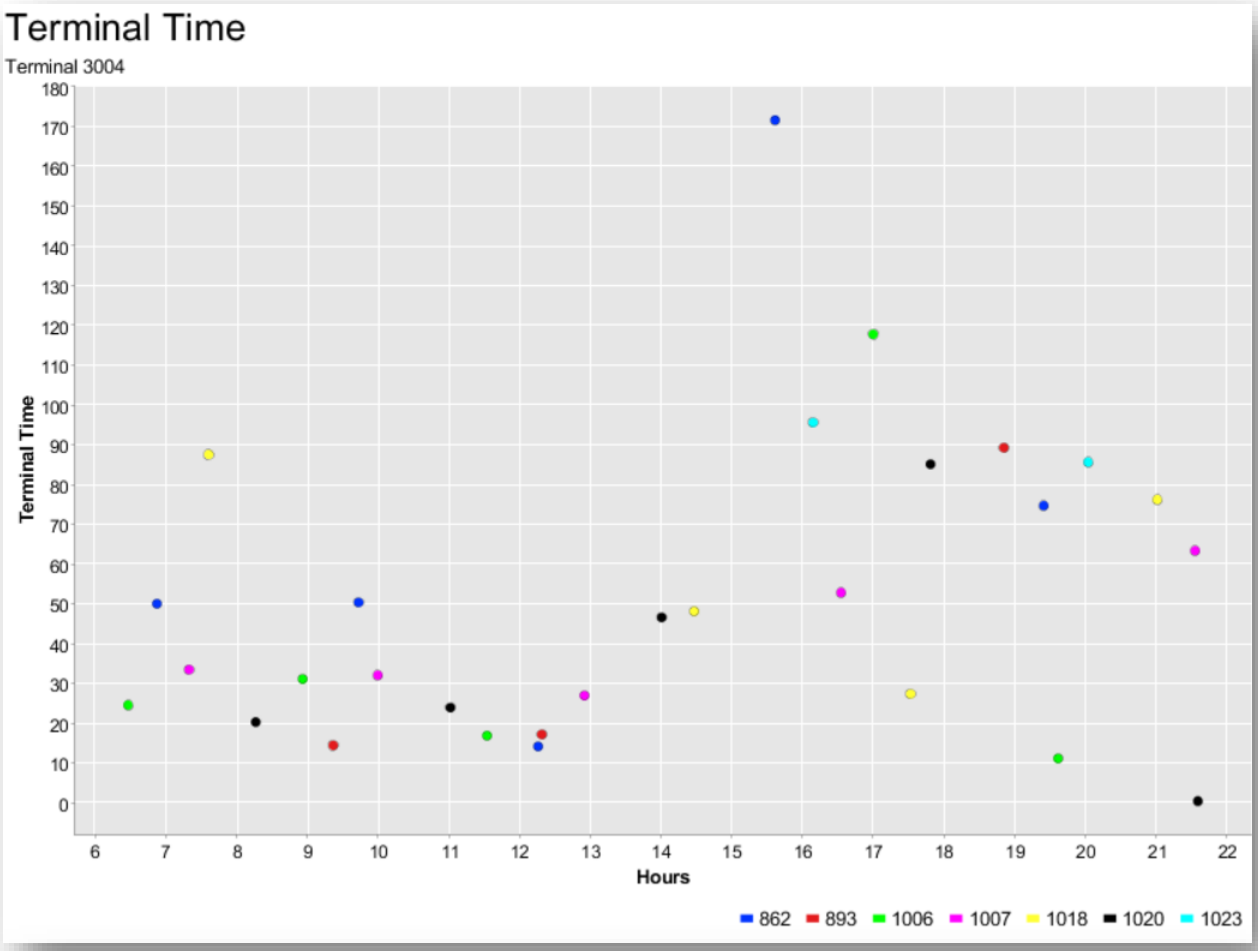
Figure 64 One-way running time 3004 to 3006



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
RunningTime	27.117	79.933	40.84	15.863	251.634	1.519	1.355	27 80

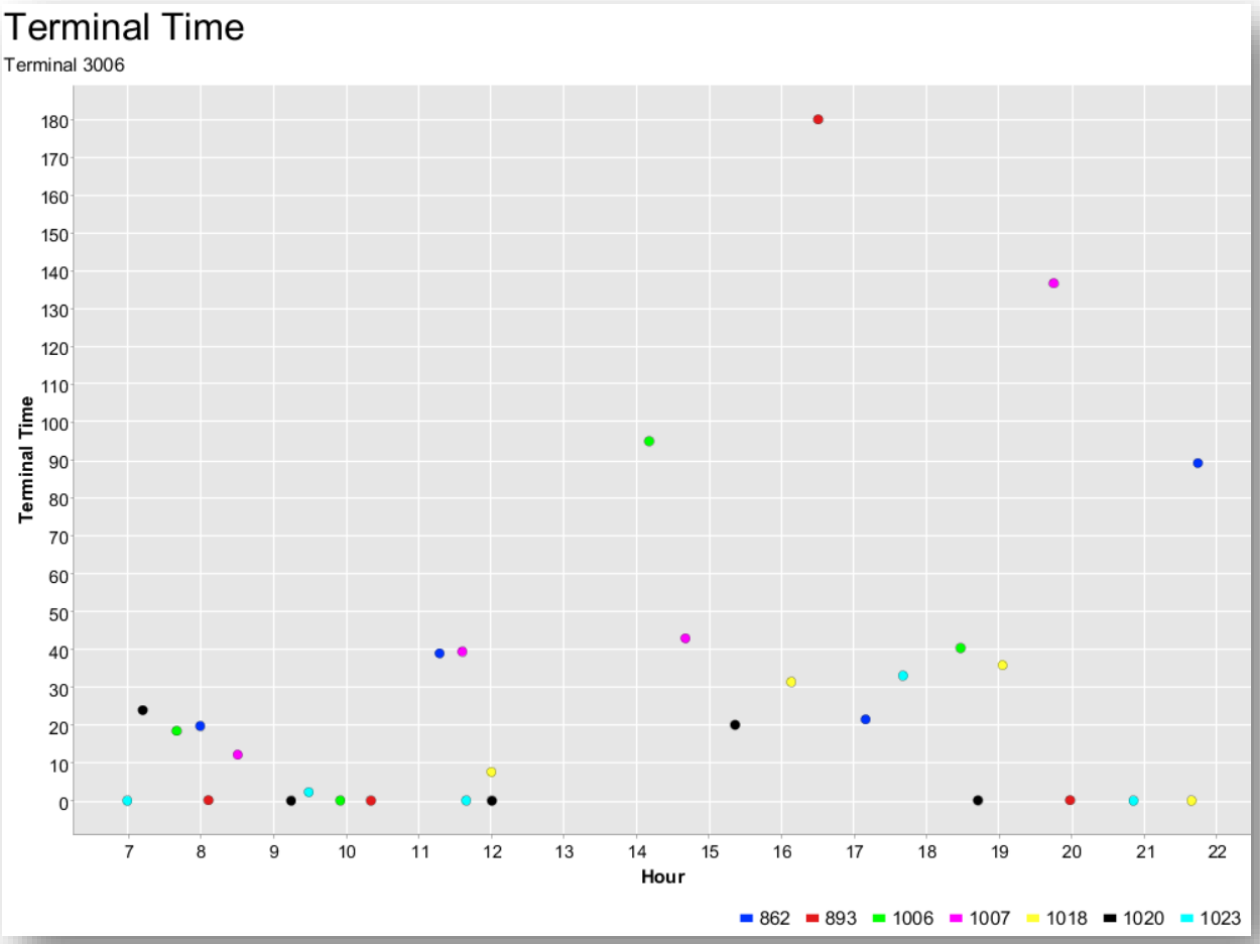
Figure 65 One-way running time 3006 to 3004

11.3.3 ONE DAY TERMINAL TIME



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
Terminal Time	0.417	117.75	47.032	31.194	973.049	0.569	-0.728	0 118

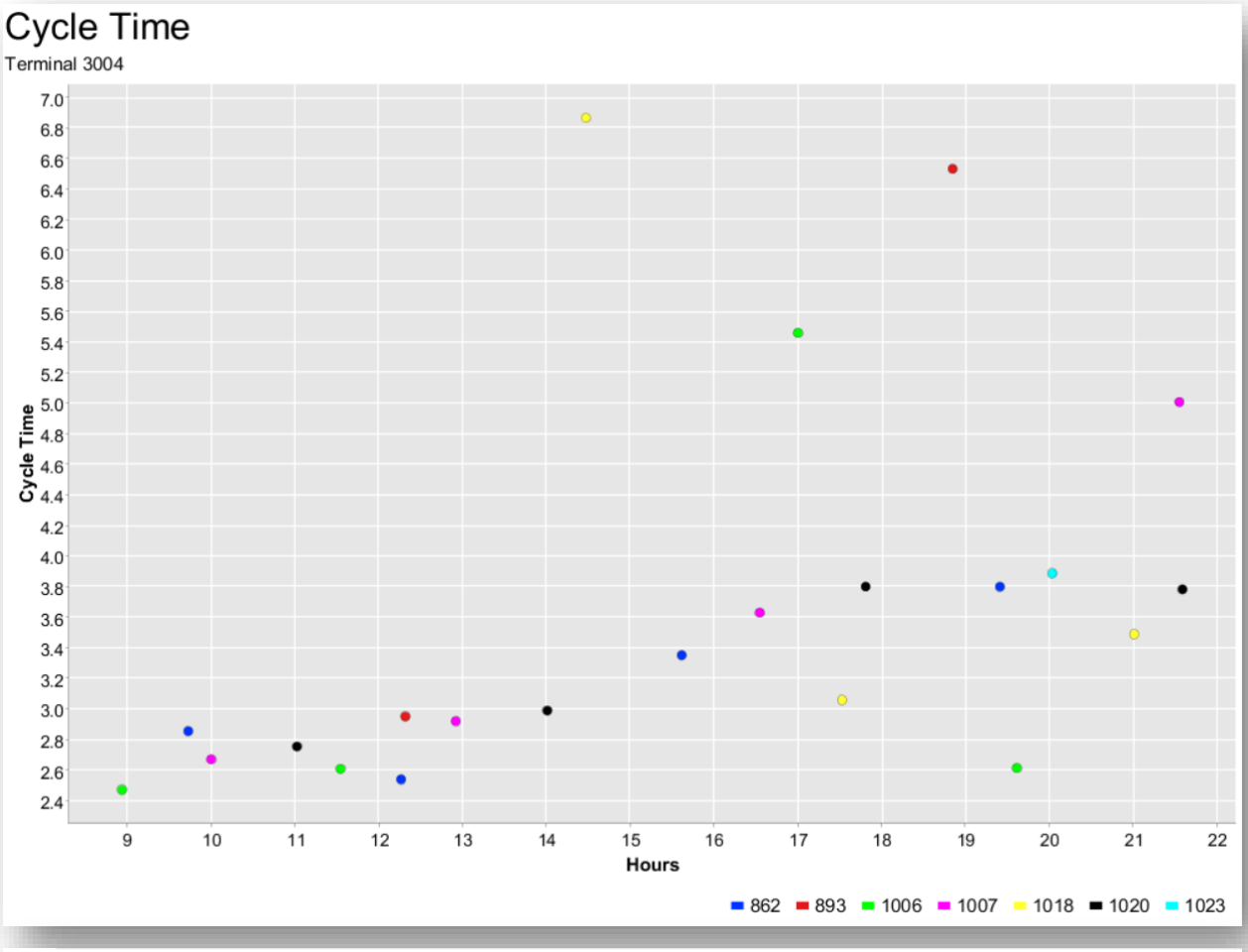
Figure 66 Terminal Time, Terminal 3004



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
Terminal Time	0.083	95.017	20.515	25.485	649.486	1.628	2.766	0 95

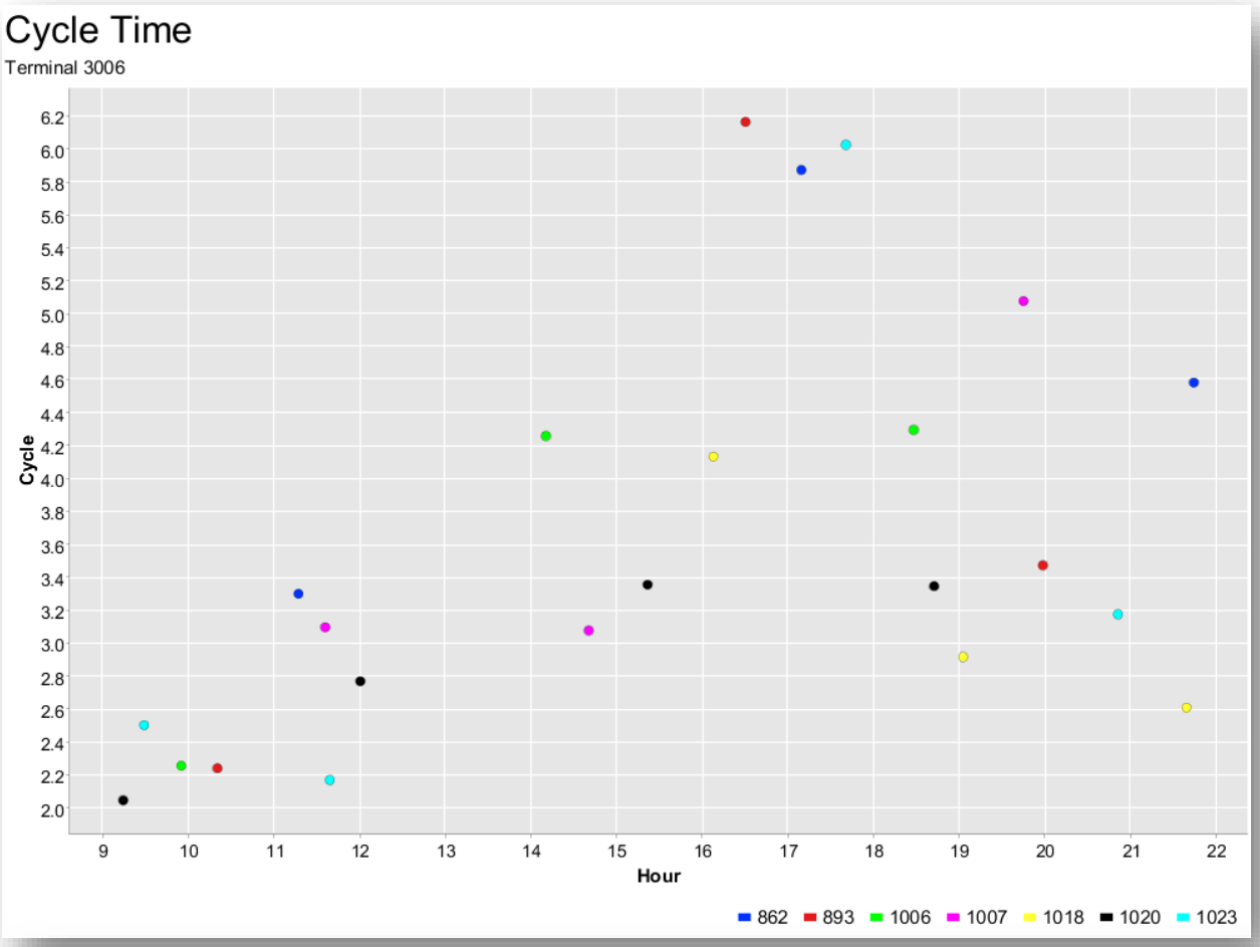
Figure 67 Terminal Time, Terminal 3006

11.3.4 ONE DAY CICLE TIME



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
Cycle Time	2.47	5.459	3.331	0.805	0.648	1.365	1.714	2 5

Figure 68 Descriptive and Scatter Plot for Cycle on Terminal 3004



Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Histogram
Cycle Time	2.045	6.026	3.48	1.144	1.31	0.896	0.124	2 6

Figure 69 Descriptive and Scatter Plot for Cycle on Terminal 3006

11.3.5 ONE DAY TRAVEL SPEED

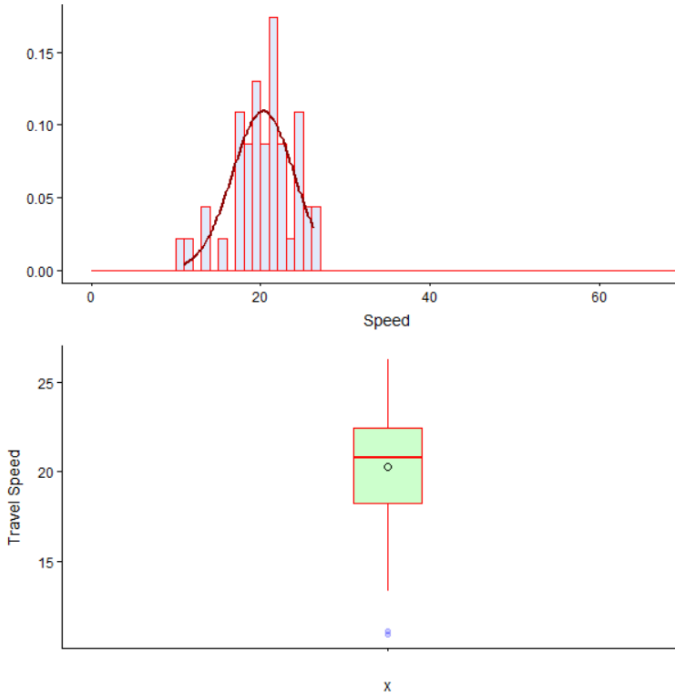


Figure 70 Histogram and Boxplot of the Mean Travel Speed of all buses from Route 5

S Column	D Min	D Max	D Mean	D Std. de...	D Variance	D Skewness	D Kurtosis
TravelSpeed	10.897	26.252	20.285	3.647	13.299	-0.646	0.528

Figure 71 Descriptive of Travel Speed of all buses from Route 5

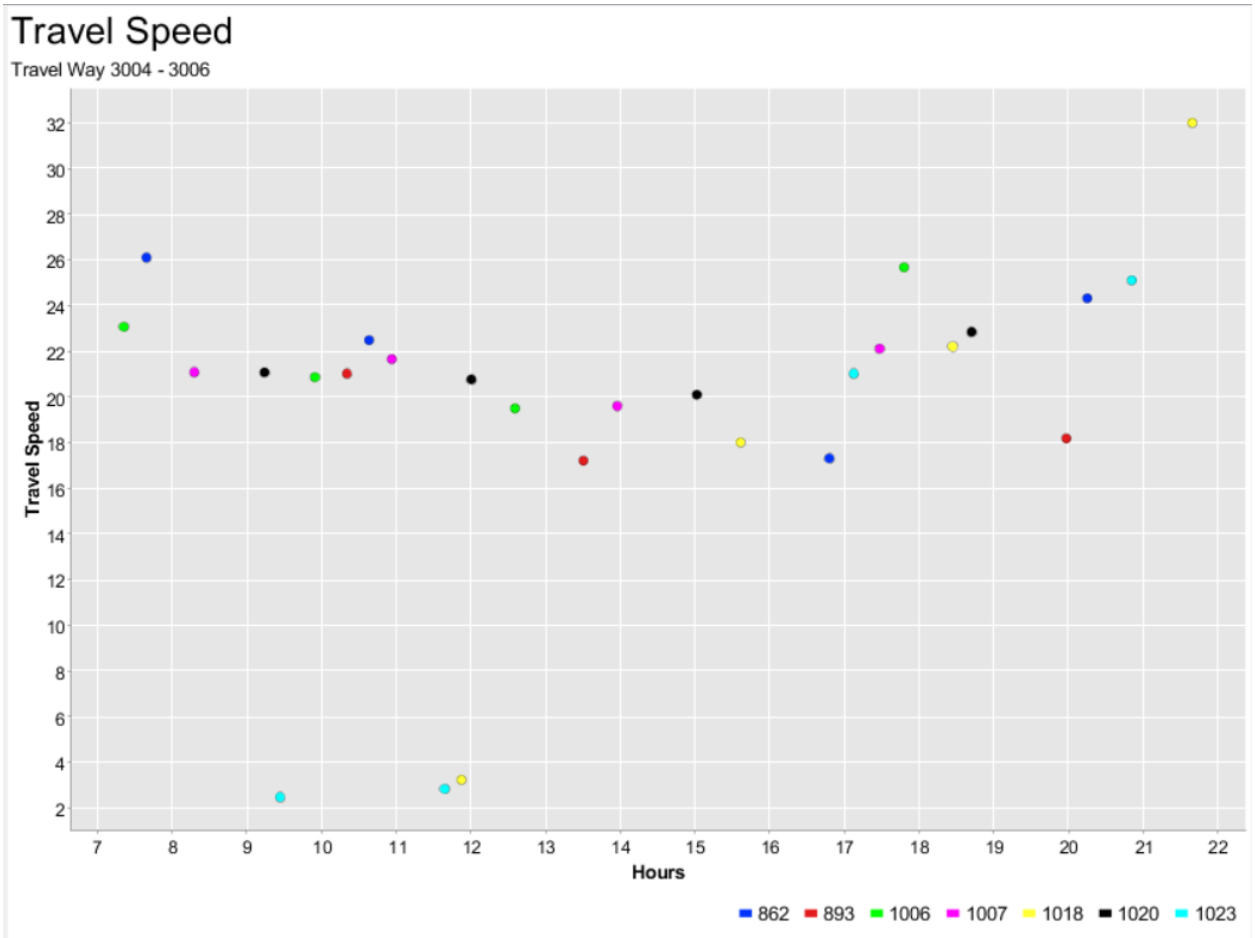


Figure 72 One-way travel speed 3004 to 3006

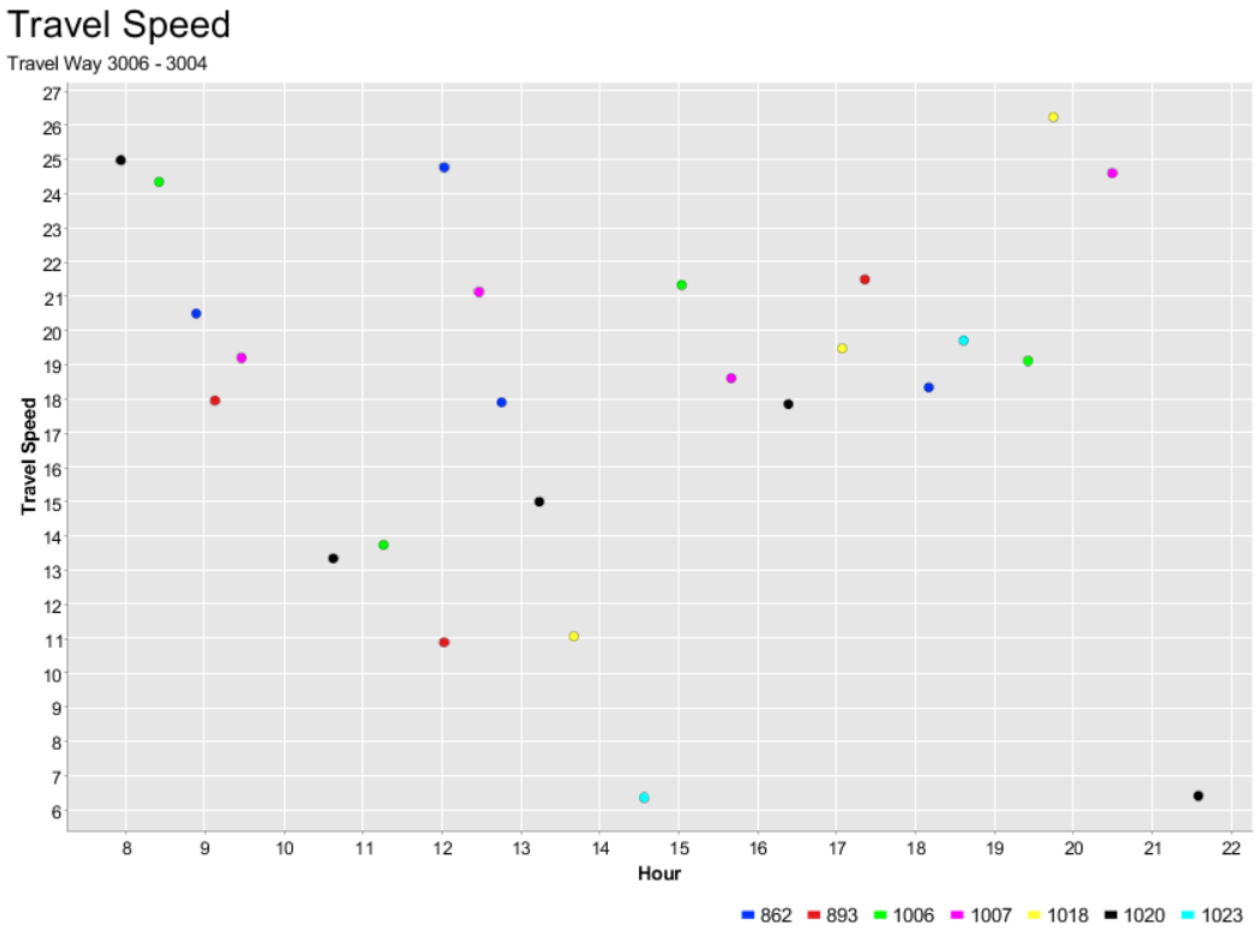


Figure 73 One-way travel speed 3006 to 3004

11.4 APPENDIX D

11.4.1 R Output for Statistical Model

```
Call:
glm.nb(formula = Boarding ~ MVR + IHS + SBI + Terminal + cIncome +
    cDensity + dMedia + Hospital + School + Governmental + Industrial +
    Recreation + Tourism, data = knime.in, init.theta = 1.175706429,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5212  -1.1904  -0.3383   0.3089   2.1085

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.000e+00  1.829e+00  -3.827 0.000130 ***
MVR           5.674e+00  2.069e+00   2.742 0.006104 **
IHS           1.253e-01  3.732e-01   0.336 0.736997
SBI           7.116e+00  1.523e+00   4.674 2.96e-06 ***
Terminal      2.649e+00  7.076e-01   3.744 0.000181 ***
cIncome      -5.721e-03  1.352e-02  -0.423 0.672143
cDensity      5.841e-05  3.710e-05   1.574 0.115447
dMedia       -4.959e-04  4.776e-04  -1.038 0.299181
Hospital      7.031e-01  2.429e-01   2.894 0.003798 **
School        3.637e-01  2.218e-01   1.640 0.101075
Governmental -6.991e-02  3.351e-01  -0.209 0.834721
Industrial   -1.962e-01  8.206e-01  -0.239 0.811019
Recreation    4.915e-01  3.494e-01   1.407 0.159500
Tourism       1.653e-02  3.089e-01   0.054 0.957322
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.1757) family taken to be 1)

    Null deviance: 218.82  on 116  degrees of freedom
Residual deviance: 132.52  on 103  degrees of freedom
AIC: 682.85
```

Figure 74 Negative Binomial Statistical Model for 04-03-2013

```
Call:
glm.nb(formula = Boarding ~ MVR + IHS + SBI + Terminal + cIncome +
    cDensity + dMedia + Hospital + School + Governmental + Industrial +
    Recreation + Tourism, data = knime.in, init.theta = 1.16319006,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4130  -1.0951  -0.3700   0.4024   2.3243

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.797e+00  1.728e+00  -1.040  0.29851
MVR           3.012e+00  1.738e+00   1.733  0.08313 .
IHS          -1.087e+00  4.998e-01  -2.174  0.02971 *
SBI           2.681e+00  1.150e+00   2.332  0.01971 *
Terminal      3.393e+00  7.209e-01   4.707 2.51e-06 ***
cIncome       3.186e-03  1.458e-02   0.218  0.82705
cDensity      5.415e-05  3.675e-05   1.474  0.14059
dMedia        6.883e-05  4.726e-04   0.146  0.88421
Hospital      7.552e-01  2.515e-01   3.003  0.00267 **
School        3.478e-01  2.245e-01   1.550  0.12124
Governmental -2.371e-02  3.414e-01  -0.069  0.94463
Industrial   -2.722e-02  7.871e-01  -0.035  0.97241
Recreation    3.678e-01  3.518e-01   1.046  0.29578
Tourism      -5.128e-02  3.128e-01  -0.164  0.86979
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.1632) family taken to be 1)

    Null deviance: 217.13  on 116  degrees of freedom
Residual deviance: 132.86  on 103  degrees of freedom
AIC: 684.11
```

Figure 75 Negative Binomial Statistical Model for 04-10-2013

```
Call:
glm.nb(formula = Boarding ~ MVR + IHS + SBI + Terminal + cIncome +
    cDensity + dMedia + Hospital + School + Governmental + Industrial +
    Recreation + Tourism, data = knime.in, init.theta = 1.148466038,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2406  -1.1312  -0.4008   0.3589   2.2193

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.535e+00  2.730e+00  -1.661 0.096645 .
MVR           3.356e+00  3.032e+00   1.107 0.268312
IHS          -2.969e-01  2.886e-01  -1.029 0.303490
SBI           7.589e+00  2.065e+00   3.675 0.000238 ***
Terminal      2.162e+00  8.907e-01   2.427 0.015226 *
cIncome      -6.966e-03  1.393e-02  -0.500 0.617059
cDensity      8.864e-05  3.701e-05   2.395 0.016612 *
dMedia       -2.099e-04  4.913e-04  -0.427 0.669248
Hospital      8.625e-01  2.570e-01   3.356 0.000790 ***
School        3.265e-01  2.242e-01   1.456 0.145266
Governmental -3.051e-01  3.470e-01  -0.879 0.379213
Industrial    1.065e-02  8.244e-01   0.013 0.989690
Recreation    4.826e-01  3.515e-01   1.373 0.169734
Tourism       1.644e-01  3.100e-01   0.530 0.595953
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.1485) family taken to be 1)

    Null deviance: 215.13  on 116  degrees of freedom
Residual deviance: 131.95  on 103  degrees of freedom
AIC: 684.29
```

Figure 76 Negative Binomial Statistical Model for 04-17-2013

```
Call:
glm.nb(formula = Boarding ~ MVR + IHS + SBI + Terminal + cIncome +
    cDensity + dMedia + Hospital + School + Governmental + Industrial +
    Recreation + Tourism, data = knime.in, init.theta = 1.136755788,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4344  -1.0453  -0.3724   0.3898   2.5471

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.473e-01  1.681e+00  -0.504 0.614222
MVR          -2.335e+00  6.492e-01  -3.596 0.000323 ***
IHS           7.430e-02  1.553e-01   0.478 0.632434
SBI           7.347e+00  3.450e+00   2.129 0.033229 *
Terminal      2.502e+00  7.669e-01   3.262 0.001107 **
cIncome       2.320e-03  1.428e-02   0.162 0.870945
cDensity      8.741e-05  3.810e-05   2.294 0.021763 *
dMedia       -7.168e-04  4.983e-04  -1.438 0.150293
Hospital      9.975e-01  2.556e-01   3.903 9.52e-05 ***
School        3.864e-01  2.274e-01   1.699 0.089250 .
Governmental -3.347e-01  3.487e-01  -0.960 0.337149
Industrial   -1.214e-01  8.025e-01  -0.151 0.879764
Recreation    4.650e-01  3.552e-01   1.309 0.190472
Tourism      -3.997e-03  3.098e-01  -0.013 0.989706
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.1368) family taken to be 1)

    Null deviance: 213.53  on 116  degrees of freedom
Residual deviance: 132.12  on 103  degrees of freedom
AIC: 685.35
```

Figure 77 Negative Binomial Statistical Model for 04-24-2013