

UNIVERSITY OF SAO PAULO  
POLYTECHNIQUE SCHOOL  
DEPARTMENT OF TRANSPORT

DOUGLAS F. W. CAPELOSSI MARTINS

A scalable method for origin-destination demand estimation using automatic vehicle  
identification data

São Paulo

2022

DOUGLAS F. W. CAPELOSSI MARTINS

**A scalable method for origin-destination demand estimation using automatic  
vehicle identification data**

**Versão Corrigida**

Dissertation presented to Polytechnic  
School of the University of São Paulo for  
the master of sciences.

Area of academic major:

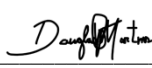
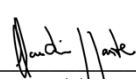
Transport Engineering - Spatial  
Information

Mentor: Prof. Dr° Claudio Luiz Marte

São Paulo

2022

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.	
São Paulo, <u>19</u> de <u>Julho</u> de <u>2022</u>	
Assinatura do autor:	<u></u>
Assinatura do orientador:	<u></u>

#### Catálogo-na-publicação

Martins, Douglas

A scalable method for origin-destination demand estimation using automatic vehicle identification data / D. Martins -- versão corr. -- São Paulo, 2022.

107 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Transportes.

1.IDENTIFICAÇÃO AUTOMÁTICA DE VEÍCULOS 2.TRANSPORTE RODOVIÁRIO 3.TRANSPORTE INTERURBANO 4.TRANSPORTE INDIVIDUAL 5.TRAJETÓRIA I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Transportes II.t.

## Resumo

A estimativa da demanda Origem-Destino (OD) é essencial para o planejamento, projeto e gerenciamento de transporte. Vários estudos recentes na literatura têm utilizado a tecnologia de Identificação Automática de Veículos (AVI) para complementar os dados tradicionais de contagem de tráfego. No entanto, extrapolar a origem e o destino das trajetórias dos veículos, combinando origem e destino para construir modelos de matrizes OD, não tem sido amplamente explorado. Portanto, este trabalho propõe um método para estimar matrizes OD para caracterizar o tráfego de veículos de passageiros, cobrindo uma malha viária de grande porte em ambiente interurbano. Os dados foram obtidos de equipamentos ITS, como contadores de tráfego e sistemas de identificação automática de veículos (AVI), na malha rodoviária do Estado de São Paulo. Utilizando sistemas automáticos de coleta de pedágios (ETC) e reconhecimento automático de placas (LPR) como fontes de dados AVI. O método teve como objetivo reconstruir uma viagem inteira com base em observações consecutivas de veículos e estimativas de zonas de origem e destino com base no modelo gravitacional clássico de distribuição de viagens. Em seguida, um algoritmo *T-Flow Fuzzy* foi aplicado para calibrar a matriz OD final. O desempenho do método foi comparado com as observações dos dados da pesquisa de campo na Região Metropolitana de São Paulo. A matriz calibrada tem um R<sup>2</sup> de 0,96, indicando que nossa metodologia fornece resultados precisos. Além disso, 65% dos pontos de contagem forneceram GEH abaixo de 5, enquanto 88% ficaram abaixo de 10, resultados considerados adequados na literatura atual. Outros estudos podem aplicar essa metodologia para analisar iniciativas de transporte público, como linhas de ônibus intermunicipais e sistemas ferroviários de passageiros, e estudos de viabilidade de concessões rodoviárias.

Palavras-chave: AVI; identificação automática do veículo; trajetória do veículo; estimativa de DO; método gravitacional.

## **Abstract**

**The Origin-Destination (OD) demand estimation is essential to transportation planning, design, and management. Several recent studies in the literature have used the Automatic Vehicle Identification (AVI) technology to complement traditional traffic count data. However, extrapolating the origin and destination of the vehicle trajectories, matching origin and destination to build matrices OD models, has not been extensively explored. Therefore, this paper proposes a method to estimate OD matrices to characterize passenger vehicles' traffic, covering a large-scale road network in an interurban environment. The data was obtained from ITS equipment, like traffic counters and AVI systems, in the Brazilian State of São Paulo road network. We used Electronic Toll Collection Systems (ETC) and License Plate Recognition (LPR) as AVI data sources. The method aimed to reconstruct an entire trip based on consecutive observations from AVIs and estimates of origin and destination zones based on the classical gravity model of travel distribution. Then, a T-Flow fuzzy algorithm was applied to calibrate the final OD matrix. The method's performance was compared to the observations of field research data in the Metropolitan Region of São Paulo. The calibrated matrix has an R2 of 0.96, indicating that our methodology provides very accurate results. Besides, 65% of the counting points provided GEH below 5, while 88% were below 10, which are considered suitable results in the current literature. Further studies could apply this methodology to analyze public transportation initiatives, such as intercity bus transit lines and passenger rail systems, and road concessions viability studies.**

Keywords: AVI; automatic vehicle identification; vehicle trajectory; OD estimation; gravitational method.

## Table of Contents

Resumo .....	3
Abstract .....	4
Table of Contents .....	5
Table of Figures .....	7
1. Introduction .....	9
1.1. Justification .....	11
1.2. Objectives .....	12
1.3. Document structure.....	13
2. Current Primary Data and Literature Review .....	15
2.1. The current state of primary data.....	15
2.2. Literature Review .....	17
2.2.1. Fixed sensor research .....	18
2.2.2. Dynamic OD matrix estimation research .....	18
2.2.3. Automatic Vehicle Identification Research .....	20
2.2.4. GPS and other on-board detectors research.....	22
2.2.5. Mobile and Big Data research .....	23
2.3. Chapter final remarks.....	24
3. Available Data .....	25
3.1. Network.....	25
3.2. Vehicle count database.....	25
3.3. Vehicle identification databases.....	26
3.4. Chapter final remarks.....	28
4. Methodology.....	29
4.1. Partial routes reconstruction (Step 1) .....	29
4.1.1. Data coverage .....	31
4.1.2. Error recognition and correction .....	32

4.1.3.	Database sampling .....	33
4.1.4.	Database simplification .....	34
4.1.5.	Partial route identification .....	34
4.2.	O/D Matrix construction (Step 2) .....	38
4.3.	Matrix calibration (Step 3) .....	43
4.3.1.	Data Input .....	44
4.3.2.	OD matrix assignment .....	44
4.3.3.	OD matrix calibration .....	45
4.3.4.	Convergence criteria.....	45
4.4.	Chapter final remarks.....	46
5.	A Method for pre-Validation .....	47
5.1.	Zone aggregation.....	47
5.2.	Assignment and R-squared analysis of model and survey data.....	53
5.3.	Chapter final remarks.....	79
6.	Results .....	81
6.1.	Partial Route reconstruction results (Step 1) .....	81
6.2.	O/D Matrix construction results (Step 2).....	84
6.3.	Matrix calibration results (Step 3) .....	87
6.3.1.	Selection of count data .....	88
6.3.2.	T-Flow Fuzzy algorithm results.....	89
6.4.	Chapter final remarks.....	91
7.	Conclusion .....	93
	References.....	94
	Glossary .....	99
	Appendix - Software and code .....	101

## Table of Figures

Figure 1: Network applied in this dissertation.....	25
Figure 2: Available data locations. ....	26
Figure 3: Potential zones for the origin-destination where data is present (red).....	28
Figure 4: Diagram of the three sections in this paper.....	29
Figure 5: Equipment coverage .....	32
Figure 6: Example of Error A.....	33
Figure 7: Visual simplified interpretation of path angles.....	35
Figure 8: A flow chart explaining the origin-destination extraction from AVI data .....	37
Figure 9: Visual representation of decimal costs on ETC equipped network links....	39
Figure 10: Visualization of the matrix estimation method.....	41
Figure 11: Result of a set of possible origin and destination zones for the selected partial route segment.....	42
Figure 12: A flow chart explaining the algorithm for extrapolating network OD and distributing trip flows under a gravitational model .....	43
Figure 13: T-Flow fuzzy diagram.....	46
Figure 14: Survey locations, <i>Pesquisa Origem-Destino 2017</i> .....	47
Figure 15: Zones, <i>Pesquisa Origem-Destino 2017</i> .....	48
Figure 16: Zones, Current Dissertation .....	49
Figure 17: SP metropolitan zone aggregation.....	50
Figure 18: <i>Baixada Santista</i> zone aggregation .....	51
Figure 19: Border aggregate regions .....	52
Figure 20: Final aggregated zones.....	53
Figure 21: State of São Paulo Metropolitan Quadrilateral.....	54
Figure 22: <i>Rodovia dos Bandeirantes</i> , trip distribution, partial routes.....	56
Figure 23: <i>Rodovia dos Bandeirantes</i> , trip distribution, zone extrapolation .....	57
Figure 24: <i>Rodovia Anhanguera</i> , trip distribution, partial routes .....	58
Figure 25: <i>Rodovia Anhanguera</i> , trip distribution, zone extrapolation.....	59
Figure 26: R-squared analysis of origin demand in outpost 809 ( <i>Anhanguera</i> ).....	60
Figure 27: R-squared analysis of destination demand in outpost 809 ( <i>Anhanguera</i> )	61
Figure 28: R-squared analysis of origin demand in outpost 810 ( <i>Bandeirantes</i> ) .....	62
Figure 29: R-squared analysis of destination demand in outpost 810 ( <i>Bandeirantes</i> ) .....	63



Figure 30: <i>Rodovia dos Imigrantes</i> , trip distribution, partial routes .....	64
Figure 31: <i>Rodovia dos Imigrantes</i> , trip distribution, zone extrapolation.....	65
Figure 32: <i>Rodovia Anchieta</i> , trip distribution, partial routes.....	66
Figure 33: <i>Rodovia Anchieta</i> , trip distribution, zone extrapolation .....	67
Figure 34: R-squared analysis of origin demand in outpost 805 ( <i>Imigrantes</i> ).....	68
Figure 35: R-squared analysis of destination demand in outpost 805 ( <i>Imigrantes</i> ) ..	69
Figure 36: R-squared analysis of origin demand in outpost 804 ( <i>Anchieta</i> ).....	70
Figure 37: R-squared analysis of destination demand in outpost 804 ( <i>Anchieta</i> ).....	71
Figure 38: <i>Rodovia Presidente Castelo Branco</i> , trip distribution, partial routes .....	72
Figure 39: <i>Rodovia Presidente Castelo Branco</i> , trip distribution, zone extrapolation	72
Figure 40: <i>Rodovia Raposo Tavares</i> , trip distribution, partial routes .....	73
Figure 41: <i>Rodovia Raposo Tavares</i> , trip distribution, zone extrapolation.....	73
Figure 42: R-squared analysis of origin demand in outpost 808 ( <i>Presidente Castelo Branco</i> ).....	74
Figure 43: R-squared analysis of destination demand in outpost 808 ( <i>Presidente Castelo Branco</i> ).....	75
Figure 44: R-squared analysis of origin demand in outpost 807 ( <i>Raposo Tavares</i> ) .	76
Figure 45: R-squared analysis of destination demand in outpost 807 ( <i>Raposo Tavares</i> ) .....	77
Figure 46: R-squared analysis of aggregate origin demand .....	78
Figure 47: R-squared analysis of aggregate destination demand.....	79
Figure 48: Partial route matrix assignment.....	82
Figure 49: Analysis of partial route matrix (Step 1) .....	82
Figure 50: Analysis of partial route matrix (Step 1) for intercity traffic.....	83
Figure 51: O/D Matrix construction (Step 2) assignment .....	84
Figure 52: Analysis of O/D Matrix construction (Step 2) .....	85
Figure 53: Analysis of O/D Matrix construction (Step 2) for intercity traffic.....	86
Figure 54: O/D Matrix construction (Step 2) without noticeable flaws.....	86
Figure 55: Intracity O/D Matrix construction (Step 2) assignment for the MRSP .....	87
Figure 56: Count data points with close proximity.....	88
Figure 57: Final count data location selected.....	89
Figure 58: R-squared of Model x Observed traffic .....	90
Figure 59: R code workflow.....	101
Figure 60: R code workflow.....	103

## 1. Introduction

Transportation engineering requires the update and development of a significant number of demand studies in which, together with other network simulation models, is essential for an extensive range of studies. They range from road concessions, public-private partnerships, reduction of greenhouse gas emissions, and improvements in transportation modes for regional and urban users.

The origin-destination (OD) matrix is essential for efficient traffic management. It designates the demand for trips between traffic zones (M. Nigro et al., 2018). Estimating the origin-destination matrix and route flows provides transportation engineers with essential data on the features of trips. Several surveys were conducted to obtain complete and quality information on the population's transportation behavior. However, in the case of Brazil, never has been such a survey applied at a state level. The biggest obstacles are the prohibitive cost and inter-regional demand for trips being lower than inter-metropolitan.

With the deployment of automated vehicle identification (AVI) systems that collect the license plates, ids, timestamps, and position of vehicles, a new dataset that can help map vehicle path trajectories were made available. For example, electronic toll collection (ETC), a typical AVI system, has been deployed in numeral high traffic density roads in Brazil.

Recently, several studies have been published where researchers use the technology to complement the compilation of traditional traffic count data, considering a complement to traditional detection, providing travel time and count data for OD estimation. Other researchers have put effort into model partial vehicle trajectory. However, extrapolating the origin and destination of the vehicle trajectories, matching origin and destination for matrices has not been extensively modeled. Another novel aspect is the scale of the study area, covering over 240,000 km<sup>2</sup>, with its challenges when it came to varying levels of population density and AVI equipment coverage. This dissertation aims to develop a method to estimate origin-destination matrices to characterize passenger vehicles' traffic in a road network covering an interurban environment with more than 1,000 traffic zones or more than 500 cities. The proposed

method uses data from ITS equipment, such as road traffic counters and vehicle identification systems (AVI), including Electronic Toll Collection Systems (ETC) on highways and radar (License Plate Recognition - LPR) in an urban environment. Additionally, socio-economic data and data from the O/D Survey of the São Paulo Metro - on the Contour Line of the Metropolitan Region of São Paulo were used.

ITS equipment, such as vehicle counter and Automated Vehicle Identification (AVI) systems, makes it possible to map vehicle trajectories from a database containing license plates, vehicle IDs, timestamps, and vehicle position. For example, Electronic Toll Collection (ETC), a typical AVI system, has been deployed in several high-traffic density roads in Brazil. In addition, studies described how the technology could complement traditional traffic data, providing travel time and count data for OD estimation (N. J. Van Der Zijpp, 1997), (M. P. Dixon et. al., 2002), (H. S. Massamani et. al., 2006). Furthermore, other researchers have tried to model partial vehicle trajectory (Y. Feng et. al., 2014), (W. Rao et. al., 2018), (E. Castillo et. al., 2008a), (E. Castillo et. al., 2008b), (C. Zhang et.al., 2019).

The method consists of three phases: path reconstruction, origin demand extrapolation - providing origin and destination selection and demand distribution - and calibration. The path reconstruction step translates AVI records into partial routes. The OD extrapolation step distributes each partial route flow into the set of most likely origin and destination zones, weighed by a combination of population and employment. The calibration step takes the estimated OD matrix and applies a fuzzy logic algorithm to adjust the matrix into matching count data.

The matrix estimation phase determines all the routes between a pair of origin and destination zones, consisting of three paths: the first stretch is from the first AVI equipment to the origin zone. The second path is a set of sections from the first AVI equipment to the last AVI equipment on which the vehicle was identified, resulting from the path reconstruction step. Furthermore, the third path is from this last AVI equipment to the destination zone. Only the routes that meet the eligibility criteria will be considered to compose the matrix, the sum of the selected routes. The matrix is the sum of trips between the origin/destination pairs. The matrix calibration phase makes

use of road traffic counters. Additionally, the validation phase uses the O/D Survey of the City of São Paulo Metro.

Commercial vehicles are subject to many additional variables such as large logistic centers, factories, ports, and farms placements. Estimating distribution weights for passenger vehicles is less affected by these variables and is represented by population and employment in a simplified way. For this reason, a decision was made to restrict the study to passenger vehicles. The increased complexity to determine distribution weights for commercial vehicles was the contributing factor. Although commercial vehicle impact on the road infrastructure is relevant to any capacity-restricted assignment model, this methodology did not apply these models in the trip distribution, as such a decision was made to not incorporate the Level of Service (LoS) impact of commercial vehicles in the network.

The contributions of this dissertation include (1) proposing a method reconstructing partial route trajectories, (2) extrapolating origin and destination through a gravity model of flow distribution, and (3) generating origin-destination matrices in a extensive scale network consisting of the State of Sao Paulo, Brazil. Additionally, (4) answers if the present AVI equipment coverage can output a matrix capable of providing a transportation profile for the region. The OD matrix outputted in this dissertation brings a more up-to-date transportation profile in the region. Furthermore, the method enables its use for datasets from 2021 forward. Further studies can utilize the OD matrix to analyze public transportation initiatives, such as inter-municipal bus transit lines, passenger rail systems, and study road concessions: expected revenue, optimal segmentation of regions achieving higher coverage, and lane expansions. Another important line is the study of congestion in intercity traffic, where lane expansions, new lanes, or new highways can be proposed.

### **1.1. Justification**

In Brazil throughout the mid-1950s, technological advancements in the automotive sector and increasing average household income made acquiring private vehicles much more accessible for most of the population. As a result, it has reduced pressure on the public transport system. However, instead of breaking this cycle, the public

policies adopted chose to encourage this demand for increased private transportation, investing more in road infrastructure to the detriment of public transit.

Presently, public transport systems require significant investment to replace private transportation, especially when discussing rail transport. When discussing this thematic at the inter-regional level, it becomes much worse since passenger railway transport in the State of Sao Paulo ceased to exist, in any significant matter, decades ago. Even in the case of freight transport, most of the rail network is inactive or idle. Existing rail concessions are only maintaining the most lucrative segments.

Although idle or in poor conditions, the lines still exist and could be reactivated by adopting proper technology and techniques. For this reason, it is necessary to carry out feasibility studies that consider the technical, economic, environmental, and financial aspects. It is important to note that these studies should be based on reliable data. Thus, this aspect will be treated with special attention in this research.

Although inferior in absolute terms, the inter-regional trip demand is one of the most significant issues in logistics and transportation in the State of Sao Paulo. The highways that approach the densely populated city of Sao Paulo are already at their limit capacity, generating high traffic and congestion. Any region with a large enough economic activity attraction is subject to inter-regional demand for trips. Although the dissertation focuses on the State of Sao Paulo, parallels with other regions in the world are conceivable.

Besides giving the possibility to estimate demand in new railway projects, Origin-destination matrices are also essential to a multitude of other transportation studies, especially on the concession of highways and infrastructure.

Given the importance of the matrices and the difficulty in obtaining them by traditional methods, this dissertation aims to use data from multiple sources of information, especially vehicle identification.

## **1.2. Objectives**

The method aims to reconstruct an entire trip based on consecutive observations from AVIs, comprising three phases. First, AVI records are converted to partial routes in a

path reconstruction. Then, we build a seed OD matrix in the second phase, extrapolating the origin from the first AVI equipment and the destination from the last AVI equipment. The ETC and LPR data processing algorithm and intermediate mapping and imaging were developed in R language, thus obtaining the seed OD matrix. Next, a fuzzy logic algorithm is used to extrapolate the initial matrix while matching to observed traffic count data.

The main objectives of this study are: 1) a method to reconstruct vehicle trajectories in a large-scale intercity road network, using ITS equipment like AVI data records; 2) estimation of origin-destination zones based on socioeconomic data; 3) seed OD matrix update using traffic count data; 4) OD matrix that adequately describes the transport profile for the Sao Paulo State region.

The proposed method was tested through an experiment that consisted of using 2017 data from 700 ETC lanes in the State of SP and over 3000 radar equipment (LPR) in urban regions, spread over an area of 248,209 km<sup>2</sup>, covering approximately 10,000 km of highways and a population of over 44 million.

### **1.3. Document structure**

The study uses approaches in different levels of detail to estimate those matrices based on automatic vehicle identification databases. The matrices resulting went through the second step of calibration, more representative of actual volume data in highways. Possible issues with the proposed methodologies are discussed, such as equipment deficiencies and biased samples. In addition, it addresses the potential of automatic vehicle identification resources to improve deficiencies with commonly constructed surveys applied to estimating and calibrating simulation models.

This paper is organized as follows. Section 2 brings the literature review, focusing on an analysis of OD estimation and trajectory reconstruction with AVI technology and looking back on previous methods using traditional count data. Section 3 outlines the available data. Section 4 describes the proposed method towards partial path reconstruction, origin and destination extrapolation, and OD matrix calibration under a fuzzy logic commercial algorithm (PTV Visum). Section 5 exposes the results and

proposed methods for validating the algorithm. Section 6 presents the conclusions drawn from the results of this study and suggests directions for further research.

## **2. Current Primary Data and Literature Review**

This chapter brings insight into the current state of primary data available to studies carried on in the State of SP. As well as a systematic review of the literature on this topic

### **2.1. The current state of primary data**

The state of SP has its secretary of transportation (STL/SP), responsible for conducting studies within the region. Their main goal is to increase their knowledge in the complex system of its many roads and transportation modes by developing and analyzing commissioned studies of various areas of expertise within transportation engineering. For this reason and to evaluate many different regions at once, in 2005, the secretary and its partner regulatory agency Artesp/SP<sup>1</sup> conducted a large-scale survey with 128 origin-destination and classified count points, distributed among 75 state roadways and three federal roadways.

In total, 114,000 interviews were conducted, over 47,000 with truck users and over 66,000 with automobile users. The survey also included more than 15,000 stated preference interviews, in 28 points distributed along with the road system divided among car (53%) and truck transportation modes.

The survey supplied input data for the generation and distribution model calibration, with cars being classified by user motive (work, leisure, other) and commercial vehicles by their number of axles (2, 3, 4, 5, 6, or more).

Many types of studies used and continue to choose matrices derived from these surveys, especially road concessions. However, despite that fact, even with this vast amount of information gathered, sampling did not capture trips from many cities, only by these trips being present outside the survey locations. For this reason, whenever a

---

<sup>1</sup> State autarchy responsible for regulating and supervising the Road Concession Program, Passenger Inter-municipal Public Transportation, and all other public transportation services delegated to the State of Sao Paulo



new study is in demand, there is still the need for the execution of complementary surveys to update the matrices.

By 2015, the secretary repeated the 2005 sample scheme (128 points) and increased it to 230 points. The purpose of sampling the 128 points from 2005 was to evaluate the changes in the decade, while new ones would aim to cover previously uncovered areas. Stated preference surveys would also expand in the amount executed.

However, a few factors would pose great difficulty for the success of this new set of surveys. Concluding these surveys would require a significant amount of necessary financial resources. Surveys also generate great disruption in traffic flow and safety concerns, and bring unidentified seasonal aspects of transportation due to short sampling periods, evidencing the main issue that this dissertation aims to propose a solution to.

The most usual method to solve this issue is conducting domiciliary surveys. The demographic Census is the most comprehensive survey available for considering every country's domicile, with its latest one concluded in 2010 by IBGE to describe the population. Additionally, it applies complete questionnaires in a Census sample evaluating population mobility.

However, its analysis does not allow the estimation of origin-destination matrices since its questions aim towards primarily migratory behavior. The focus of the Census is not to adequately evaluate trip behavior.

Due to this reason, there is a need to plan specific surveys. These domiciliary origin-demand surveys are designed to investigate trip characteristics of all domicile residents, determining destinations, frequencies, motives, modal choice, and other aspects. In addition, road and public transport (trains, airports, buses, subway) surveys on the outer edges complement this method.

The Domiciliary origin-destination survey's primary goal is to create a passenger transport model. In the State of SP, examples are:

- Metrô (Subway) and the STM (Secretary of Metropolitan Transportation) domiciliary origin-destination survey, carried on in the metropolitan region of Sao Paulo every 10 years (since 1977). The 2017 edition is in its final stages.
- STM domiciliary origin-destination survey carried on in the metropolitan region of Campinas in 2011.

## **2.2. Literature Review**

The present dissertation acquires data from the latest available technology systems in transportation engineering in the State of SP, being at the forefront of applying data from these systems in transportation models. While recent in Brazil, these systems have already been operational for more extended periods in more developed countries. Therefore, this section's purpose is to present standard practices, methods, and knowledge acquired from employing data from these sources in this present dissertation.

Although the classical 4-step model (ORTÚZAR & WILLUMSEN, 1990) is still the central line adopted by most transportation engineering companies, government agencies, and professionals, newer techniques come to supply several shortcomings due to non-computational data in previous methods concerning passenger or cargo trips. Nigro et al. (2018) state that to conduct such traffic analysis both for static and dynamic studies, specialists must have quality in: a representation of the traffic network and a data set to simulate network routes and predict the traffic flows.

An accurate and reliable OD matrix, as well as vehicle trajectory data are required for efficient urban traffic management. Conventionally, OD matrices were derived from surveys, a process that is time-consuming and usually does not reach precision enough due to biased response and reduced sample size. With the progress of traffic detection technology, newer methods were propositioned to estimate OD's matrix using different data sources. These different approaches, for the most part, can be divided into two distinct categories, the fixed-sensor-based and the trajectory-based methods.

### **2.2.1. Fixed sensor research**

Fixed sensors include radar detectors, loop detectors, and video sensors. These collect traffic information like volume, occupancy, and speed at fixed locations. Substantial research has been devoted to estimating the time-dependent traffic states such as OD demand with fixed-sensor data. Initially, Willis and May (1981) proposed a proportional distribution method. The method considered that the OD volume on roads proportional to the volume at entrance ramps. Nihan (1982) proposed a gravity-based model. The author assumed that trip distances meet a Gamma distribution, inferring that demand on both costs extremes is a slight possibility. Van Zuylen and Willumsen (1980) used the entropy minimizing principle to construct the OD estimation problem based on link volume observation. Michael (1991) adopted the generalized least squares approach to estimate OD matrices with a combination of survey and traffic count data. Yang et al. (1994) examined the problem of estimating OD matrices from traffic counts in congested networks using a least-squares technique. Since fixed sensors are not able to capture traffic's origin and destination, these aforementioned methods introduce assumptions to estimate an OD matrix.

### **2.2.2. Dynamic OD matrix estimation research**

As for dynamic OD matrix estimation, Cremer and Keller (1987) developed four approaches to identify OD flows and tested their performance using synthetic and actual data from several intersections. The results show greater accuracy and demonstrated the advantage of the dynamic approaches to the static estimation procedures.

Lin and Chang (2007) applied estimated travel time distributions to OD matrix estimation. Sherali and Park (2001) proposed a parametric optimization based on a least-squares model to determine time-dependent trip tables. Furthermore, the algorithm needed additional time-dependent shortest-path subproblems to solve the problem, generating additional path information. The published methodology was appropriate only for offline processing purposes. Xie et al. (2011) proposed a maximum entropy model (ME-LS) estimator for elastic OD flow tables and tested its results on the Sioux Falls sample network. Castillo et al. (2014) give a Bayesian statistical

approach, to study the gamma-based hierarchical optimization problem to estimate OD matrices. They proposed a multi-level approach consisting of: (1) a Wardrop minimum variance assignment model for deriving the route choice probabilities, (2) a least-squares problem for obtaining OD sample data, and (3) a maximum likelihood problem aimed at estimating the posterior modes. They applied the method to a sample network and a medium-sized city (Ciudad Real). The proposed method seems to be sufficiently validated, providing similar flows to existing techniques.

Tobias and Bernhard (2013) proposed a combined method for short-term detector forecasting in urban locations and traffic demand estimation using the forecasted counts as constraints for estimating OD flows, route, and link volumes. Jiang et al. (2011) and Mussone et al. (2010) employed the neural network for large-scale OD estimation. They concluded that the developed methods enable capturing the spatial-temporal correlations between OD demands. Lee et al. (2011) presented a dynamic OD estimation model based on a three-phase traffic theory. Real-time traffic data, such as traffic flows, speed, and occupancy, was used to estimate the dynamic OD demand between the on-ramp and off-ramp on the freeways. Perakkis et al. (2012) applied a Bayesian statistical approach to incorporate trip-generation, trip-attraction, and trip distribution in one model. A model of OD flows derived from census data associated to a set of explanatory variables is presented.

Another source of fixed data source comes from Bluetooth and Wi-Fi. Bugeda et al. (2010) simulated an experiment before deploying AVI technologies by emulating the logging and time stamping of a set of vehicles equipped with Bluetooth and Wi-Fi mobile devices. The detection of these devices could provide estimates of travel times and OD patterns for the entire population of vehicles, and *ad hoc* procedures based on Kalman filtering were successfully implemented. Barcelo et al. (2010) looked into the quality of the data produced by Bluetooth detection of mobile device equipped vehicles for travel time forecasting and developed a Kalman Filter method to estimate time-dependent OD matrices in highways. Suitable results were achieved in uncongested traffic conditions.

### 2.2.3. Automatic Vehicle Identification Research

In contrast to traffic detector data typically used by the previous researchers, automatic license plate recognition data (AVI) represent another important and emerging data source for estimating dynamic OD demands and serving traffic network management. The trajectory data records the movement of distinct vehicles, therefore providing a consistent data source for OD demand estimation. Unlike traditional fixed sensors, they identify each vehicle or traveler via an identification (ID), which makes trajectory reconstruction possible. Van der Zijpp (1997) proposed a constrained optimization formulation to estimate OD demand and identification rates together with license-plate-based AVI data. Dixon and Rilett (2002) calculated link flow proportions based on observed travel time from AVI counts. They presented offline generalized least squares and online Kalman filtering models for estimating OD demand.

Working with data collected near the Olympic Park in Beijing, China, Feng et al. (2015) applied particle filter theory combining five spatial-temporal trajectory correction factors to estimate the vehicle's trajectory. The proposed method demonstrated high accuracy (90%) for reconstructing trajectories when AVI coverage is 50%. Zhou and Mahmassani (2006) proposed a nonlinear ordinary least-squares model. Using a simplified Irvine, California testbed network, they combined AVI counts with other available information sources into a multi-objective optimization framework and exploited OD demand distribution information based on synthetic data.

Rao et al. (2018) introduced an offline method for historical OD pattern estimation based on AVI data. First, a particle filter model. By searching potential paths in pre-determined areas based on time geography theory, it was possible to generate the initial particles. Through the reconstruction of completed trajectories of all vehicles in numerous trips, path flow estimation is determined. The study also shows a minimum AVI sampling rate (60% for their Kunshan, China network) for estimating the OD patterns with reasonable accuracy. Castillo et al. (2010) discussed the problem with optimizing the usage of scanning technology for traffic estimation, mainly route flow. Considering three problems: minimizing the usage of resources used to estimate a given subset of flows, identifying the selection of scanned segments for a pre-determined number of cameras, and solving the previous problems considering errors

in scanning and error recovery. First, an optimization problem was solved with the CPLEX solver in General Algebraic Modeling System (GAMS). The method is then applied in a medium-sized and straightforward network of Cuenca.

Fu et al. (2017) proposed a method using stochastic integer programming and branch and bound integer L-shaped algorithms. The first stage they approached minimizing total traffic scanning installation cost and the impact of paths not covered. Their second stage attempted to model and reach a solution that minimizes paths not covered for a given scenario and sensor locations. Castillo et al. (2008a) attempted to reconstruct the path flow to estimate the OD matrix through a Bayesian network and the Wardrop minimum variation model. He also examined the impact of the layout of AVI facilities on OD estimation based on the path-flow reconstruction method (2008b).

More recent research regarding optimal placement and locating scanning devices are present in Sanchez Cambronero et al. (2020). For obtaining the necessary data for analyzing traffic and making network forecasts, the authors aimed to address the fact that current methodologies aimed at network modeling and data processing are not fully adapted for the usage of license recognition devices. Route flows are an essential variable in models that used data from plate scanning (predominantly AVI sensors). At the same time, traditional methods are based on observing link and/or OD flows.

On the subject of historic automatic fare collection (AFC) data usage, Yang et al. (2020) introduce a nonlinear programming model for predicting the dynamic OD matrix for an urban rail transit system. The model assigns passenger flow to the hierarchical flow network, calibrated by backward propagation (BP) of the first-order gradients and reassignment of the passenger flow with the benefit of assigning updated weights between different layers. Zhang et al. (2019) extracted OD patterns with historical trajectory data to simulate Ramp Metering as an effective measure to alleviate freeway congestion. The research shows that ramp metering with trajectory data increases the throughput by another 4% compared with conventional fixed-sensor data, displaying a significant advantage under heavy traffic, situations in which traditional control loses effectiveness.

In order to obtain the OD matrix and the traffic volume metrics, Teknomo et al. (2012) converted trajectory data into a group of linear algebraic equations to represent the relationship between the OD matrix and the path flow. Parry et al. (2012) integrated discrete trajectory and traffic volume data to analyze OD estimation based on the maximum likelihood estimation method. Using high-accuracy electronic toll data, Kwon (2006) used a simple moment estimation method based on electronic vehicle tags for dynamic OD estimation. Kwon et al. (1994) used sampled vehicle trajectories to estimate a time-dependent OD trip table, with the addition of restrictive assumptions for a complete traffic assignment map.

#### **2.2.4. GPS and other on-board detectors research**

On-board detectors return yet another data source on vehicles. With detailed traffic data collected in the Chengdu, China city center, Ásmundsdóttir (2008) constructed matrices, analyzed route choices and trip lengths. The data consists of traffic counts, video camera data, and taxi floating car data (FCD). He concluded that FCD can be employed for estimating matrices and analyzing the route choices. However, noted that FCD data lack some information, due to the fact of being only sample data.

Yang et al. (2017) presented two OD estimation models using sampled GPS position data of probe vehicles and link flow counts: (1) scaled probe OD as prior OD (SPP), and (2) probe ration assignment (PRA). The SPP model uses scaled probe vehicle matrices as prior matrices and applies conventional generalized least squares (GLS) into bringing OD correction with link counts. The second model (PRA) is an extension of SPP with observed link ratios as additional information in the estimation procedure. Under the circumstance of heterogeneity of probe penetration ratios among different OD pairs, the PRA model would outperform SPP. Such a situation could occur principally when probe vehicles are a specific type of commercial vehicle. Huang et al. (2018) applied the human mobility model estimating hourly travel demands for Shenzhen, China. He proposed a model combining the advantages of mobile phone data with urban transportation data to predict crowd gatherings that commonly originate traffic jams.

Other authors introduced GPS data, such as Ibarra-Espinosa et al. (2019) and Moreira-Matias et al. (2016). However, because not every vehicle is equipped with GPS devices or not all travelers use navigation apps, Eisenman et al. (2004) presented that probe penetration rate influences estimation accuracy. His research also noted that adding probes has a significant value in providing estimates for OD flows. By introducing a small percentage of probe trips (e.g. 10%), when no prior seed matrix is introduced, improvement of matrix estimates by more than two order of magnitudes is possible.

### **2.2.5. Mobile and Big Data research**

Herrera et al. (2008) propose and evaluate two approaches to reconstruct path flow by employing mobile data and data collected by stationary detectors. The first approach is based on data assimilation methods (so-called Nudging method or Newtonian), and the second is based on Kalman filtering.

Toole et al. (2005) use call detail records (CDRs) from mobile devices in association with open and crowd-sourced census records, geospatial data, and surveys. Daily trips were constructed through an analysis of consecutive observations from users at different stop points during determined time frames. Zin et al. (2018) continues with the usage of CDR in Yangon, the economic center of Myanmar. Yang et al. (2020) pointed out that with traffic flow estimation, classical statistical methods are still widely applied in not only short-term predictions, but for more generalized studies as well. Nevertheless, machine learning methods are also shown to be very useful due to their many advantages, for example, problem adaptability, generalization, and learning ability, which is very important to estimate traffic flows using field data. For example, Sanchez Cambronero et al. (2010) used Bayesian networks, Bai and Chen (2019) used neural networks, and Lui et al. (2018) used deep learning.

By combining research findings, partial vehicle trajectory represents a new approach to solving travel time analysis, construction relationships between path and links, path flow estimation, and OD demand acquisition. Research on reconstructing a complete vehicle trajectory based on a partial trajectory and corresponding spatial-temporal data is continuously improving.



### **2.3. Chapter final remarks**

This chapter discussed the current stage of primary data and its many deficiencies and exposed the most recent attempts to collect information essential for transportation forecasting.

In sequence, this dissertation analyzed a selection of the most recent and relevant studies on utilizing AVI data for demand estimation.

There have been many successful methods that use AVI datasets in generating OD demand estimations, some more complex and comprehensive than the method proposed in this dissertation. However, these studies all focused on smaller networks and highly sampled information, failing to fulfill the requirements of country-wise analysis, with its large networks and vastly more significant amounts of OD pairs. This dissertation uses a different approach with a few aspects from other studies. Such as gravitational attraction model (in contrast with particle filtering) and consecutive observations from users at different points, more familiar with mobile device data, to estimate paths on the network-defined OD zones.

### 3. Available Data

In this dissertation, two separate databases are used as primary data sources. One being the vehicle count database (VCD) and the other being the Vehicle Identification Database (VID). Secondary information originates from other data sources, primarily economic and demographic information.

#### 3.1. Network

The network used in this dissertation (shown in Figure 1) is a product of the work completed by the Logistics and Transportation Secretary and is under its usage policy.

Figure 1: Network applied in this dissertation



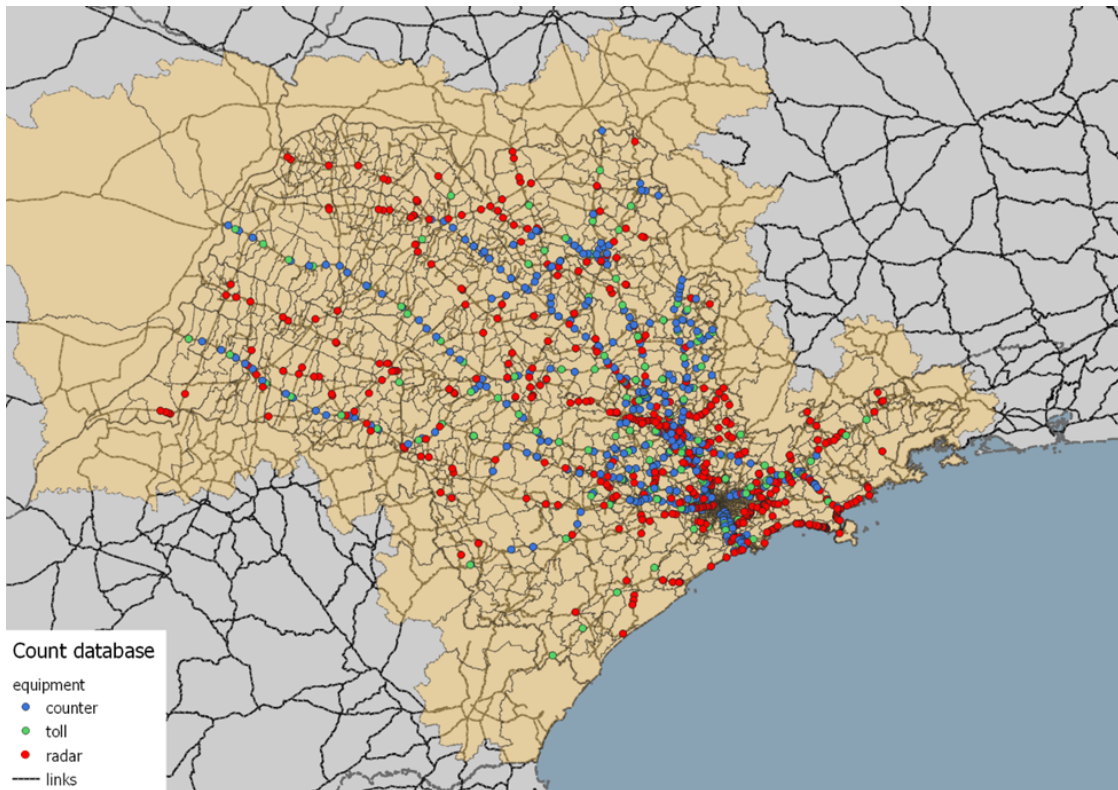
Source: Author

#### 3.2. Vehicle count database

The vehicle count database (VCD) is organized by a sequence of volumetric data collected by toll plazas, SAT (traffic analysis system) equipment, and radars. Its

information spans a broad coverage of state road infrastructure Figure 2 shows the available data locations.

Figure 2: Available data locations.



Source: Author

This dissertation was given access to 1,192 count data points originating from the state (Artesp/SP). Caution is necessary with the usage of a high amount of collection points. While it increases the calibration and validation of the model, it could also bring conflicting information that makes the calibration process incapable of reaching a balance. A method of filtering the desired set of count locations is discussed in item 5.3.1.

### 3.3. Vehicle identification databases

In 2018, the Logistics and Transportation Secretary started to organize, tabulate, and store databases that come from systems that use different technologies capable of registering vehicles.

Plate (or another ID) registration, date, and time of passage of vehicles, combined with the location of the specific equipment that registers the information, allows the identification of the route (partial routes) used by each trip. One of the following systems should capture vehicles plates passing through their locations:

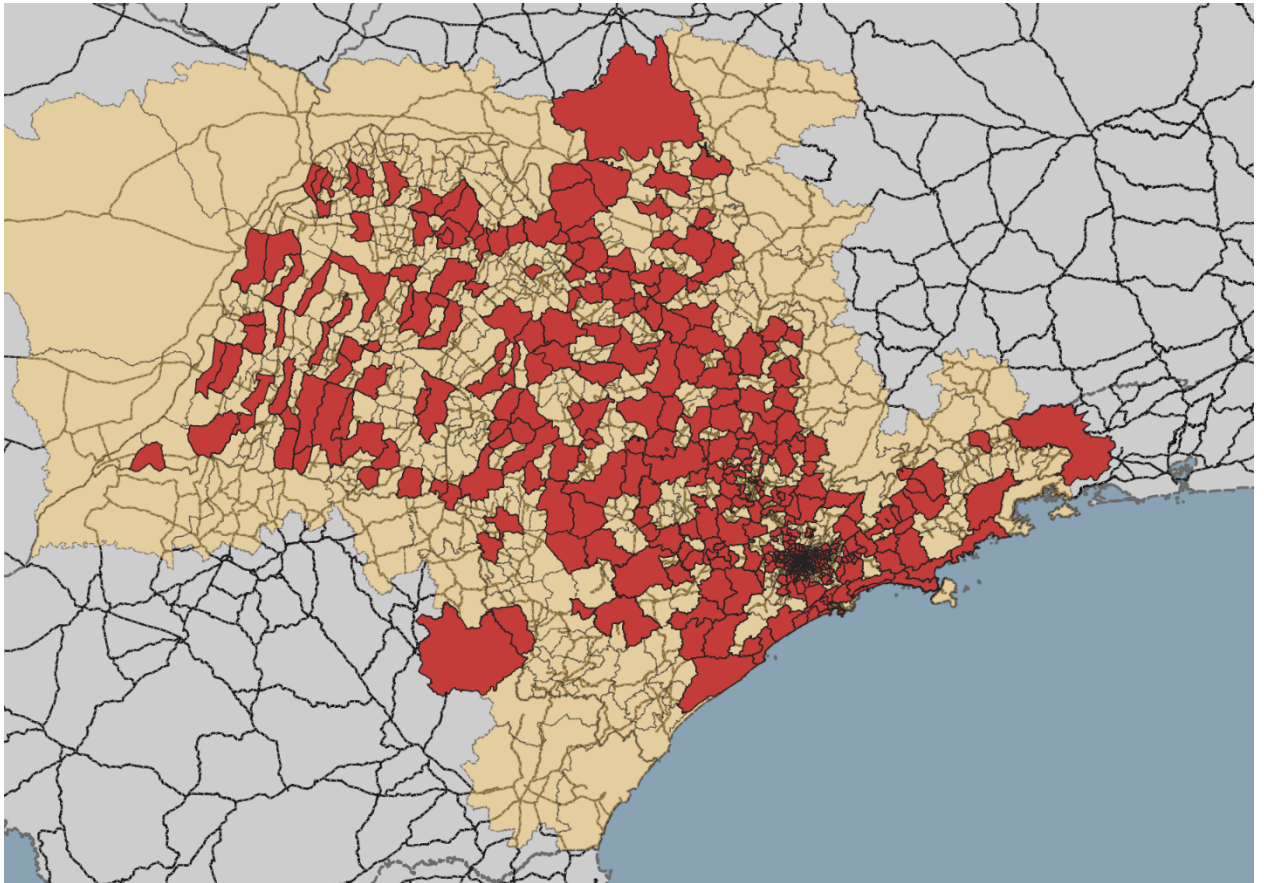
- Electronic Roadway Monitoring from DER/SP<sup>2</sup>, by the usage of OCR;
- Electronic Payment Systems by AVI, used by concessions across the State of SP (Artesp/SP);
- Monitoring Systems partnered with the DETECTA system from the Military Police of the State of SP, used by the Operations Center of the Military Police (COPOM).

Equipment present in these systems has not yet reached an outstanding coverage of the region, although it has a presence in the most highly-dense locations. Figure 3 shows the zones with at least one piece of equipment present. These systems, when combined, are configured as a central data resource. Path reconstruction methods (section 4.1) make it possible to determine a set of trips from consecutive observations.

---

<sup>2</sup> State government department with the purpose to administrate the state owned roadway infrastructure.

Figure 3: Potential zones for the origin-destination where data is present (red)



Source: Author

### 3.4. Chapter final remarks

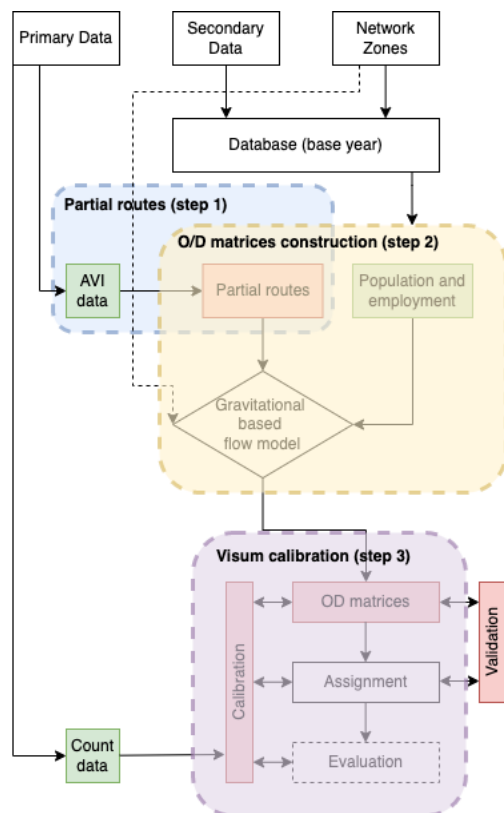
The corresponding section gives an overview of the data available to the dissertation and its potential for, alongside the method proposed in the next chapter, fulfilling the primary goal of this dissertation. The section brings insight into the deficiencies in the data and the strategy of the method proposed to circumvent them.

## 4. Methodology

This chapter discusses and presents the proposed methods for partial route reconstruction, OD extrapolation and distribution, and OD matrix calibration. The adopted strategy after the partial route's reconstruction was to investigate the expansion algorithm on each of the main São Paulo highways that connect the RMSP to the rest of the state. This strategy configures a pre-validation step before the matrix calibration.

The method is split into three sections, each providing valuable information to the following step. Figure 4 shows each step and how they tie in together.

Figure 4: Diagram of the three sections in this paper



Source: Author

### 4.1. Partial routes reconstruction (Step 1)

This step is designed to translate the vehicle identification database into partial routes that consist of sequential equipment locations that correctly identify a vehicle. Data

records present in the database consist of records as presented in Table 1 and Table 2

Table 1: Structure of vehicle database – DETECTA and DER system

X coordinate	Y coordinate	Vehicle Plate	Timestamp
--------------	--------------	---------------	-----------

Source: Author

Table 2: Structure of vehicle database – AVI electronic payment system

X coordinate	Y coordinate	Vehicle Plate	Timestamp	Vehicle Category
--------------	--------------	---------------	-----------	------------------

Source: Author

These datasets differ in their capture system, data coming from DETECTA and DER systems use OCR speed scanners that record every plate captured by its software and image processing capabilities. Information captured is influenced by many variables, such as software reliability, climate variability, and readability of vehicle plates. The presence of so many variables gives each piece of equipment a potential margin of error, either by mistranslating a vehicle plate (error type A) or failing to register it (error type B). Comparing each equipment's total data entries to other types of information sources, such as vehicle counts, gives information about the order of magnitude and the prevalence of error type B. Not every scanner has other alternate equipment directly over or close to it that outputs information. This dissertation does not cover the effects of error type B and its prevalence. Bernardi (2017) discusses issues with OCR systems and brings a comprehensive insight into the technology.

The capture systems used in the AVI system do not require image capture and recognition since the tolls scan each vehicle equipped with an information microchip across the many toll plazas in the State of São Paulo. This fact results in a system with significantly higher accuracy than OCR scanners, the downside being that not every vehicle carries these microchips. Adoption of this type of technology is different across regions; rural regions show a weaker adoption of the equipment. Highly urbanized regions show high adoption percentages. According to the data available to this dissertation, the average adoption in Sao Paulo state is at 57%<sup>3</sup> by more recent

---

<sup>3</sup> <https://estradas.com.br/artesp-autoriza-nova-operadora-de-pedagio-eletronico-nas-rodovias-paulistas/>

estimates; according to the data available to this dissertation, adoption averages at 56%. The different accuracy of both systems plays an essential role in their usability. Therefore, data coming from each database are treated differently.

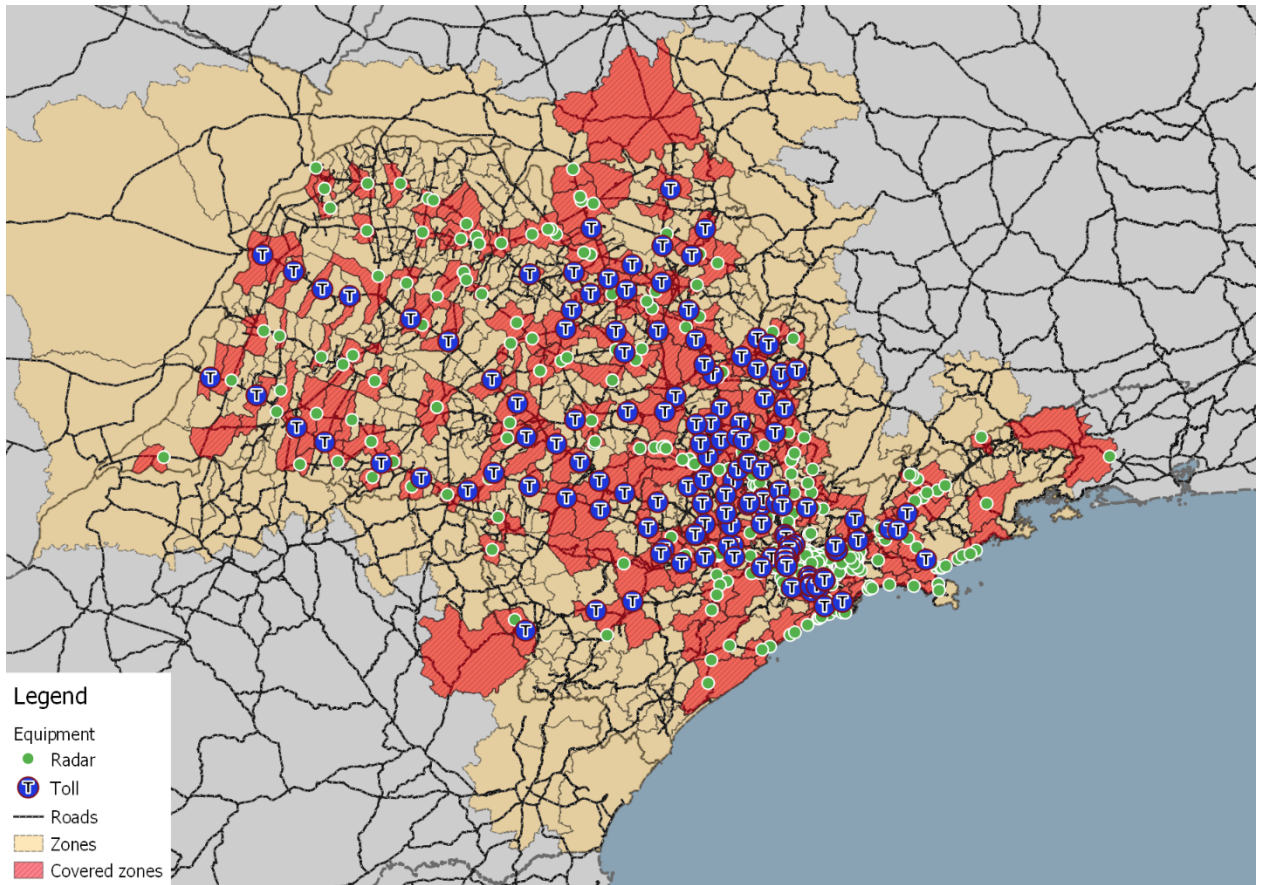
#### **4.1.1. Data coverage**

Ideally, as much granularity as possible of information is desirable. For example, having identification systems on every corner or kilometer of the road network would result in a much more reliable result of each vehicle trip's origin, destination, and midpoints.

However, as seen in the Available Data section and Figure 5, sections of the State of SP where data granularity is not ideal, showing few equipment placements through the road network, which results in lower reliability on the model at those locations. However, these regions are usually less densely populated and have a lower demand for transportation infrastructure, so that a demand estimation model could fill these voids with sufficient data. As justified before, this added model is not used in this dissertation.



Figure 5: Equipment coverage



Source: Author

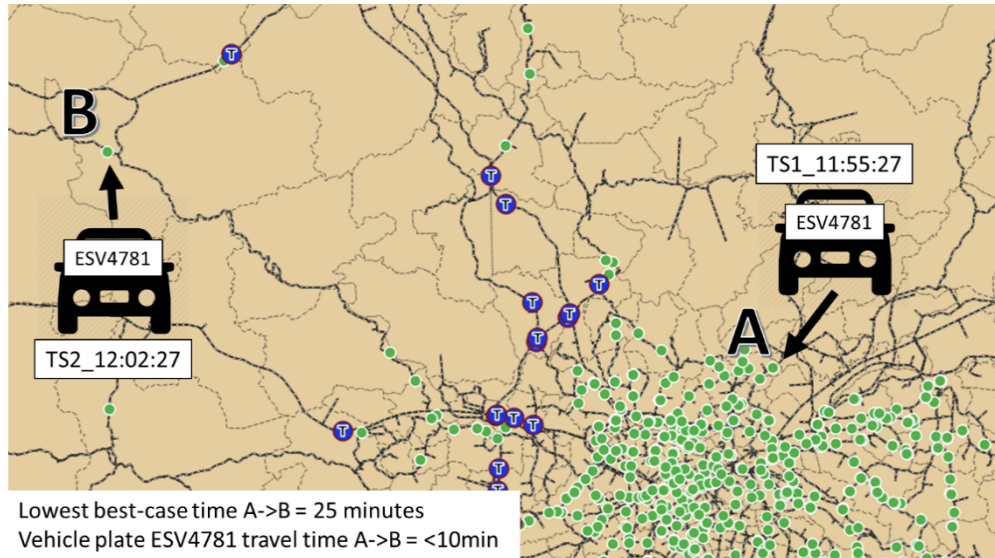
#### 4.1.2. Error recognition and correction

This step is necessary because, while automated systems have much higher data processing capabilities than manual processes, there are still errors coming from OCR software that could construct incompatible partial routes.

It is not in the scope of this dissertation to treat error type B occurrences, but for error type A, a system of verification is simple and effective to filter most of these types of errors. The method consists of for each vehicle identified in a specific timestamp, querying its previous appearances and matching the displacement in a location with what should be a sufficient amount of time for it. Any vehicle showing a length of time lower than best-case intervals for each location pair is more than likely a different wrongly identified vehicle plate by the OCR software. A visual explanation of this process is present in Figure 6. For this explanation, fictional vehicle ESV4781 passed through segment A at 11:55:27, then again at segment B at 12:02:27, giving it an 10

minutes trip time. The best-case scenario is a vehicle going from point A to B in around 25 minutes, giving an inconsistency on its speed detection of vehicle ESV4781, indicating an issue with identifying its license plate at location B.

Figure 6: Example of Error A



Source: Author

The percentage of data removed through this filtering process was less than 1% of total records.

#### 4.1.3. Database sampling

Throughout the activities in Group B, it will come of significant importance for vehicles to be present in both the DETECTA and the AVI system. The AVI system has a very high detection rate (over 99.9%), and it is safe to say that if a vehicle goes through an AVI-equipped tool, the record will show in the database. In such a manner, this step filters out every vehicle plate present in the DETECTA/DER system but not in the AVI system, with an additional 60% of records removed from the database due to this step.

#### **4.1.4. Database simplification**

The modeling network plays a crucial role in many activities across this dissertation. Nonetheless, it is still a simplified representation of the complex road network in the State of SP. Many streets and less meaningful road connections are not represented in small cities and sparsely populated regions. For this reason, many scanners coming from the DETECTA/DER database are located in uncovered points by the simulation network.

By the assignment step, some data could not be possible to model. Road networks usually have some sort of simplification proportional to their scale. In this case, equipment data outside of the road network model could not translate into accurate route information. Here, one simplification would be to group every scanner outside of the network to each zone within its borders. Then, every unique vehicle that remained within this equipment group had only its latest record kept. Grouping the scanners eliminate partial routes that would eventually become intra-zone trips.

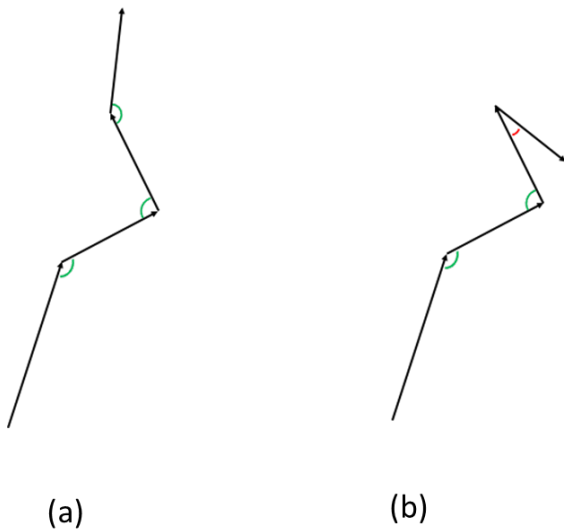
#### **4.1.5. Partial route identification**

This step is designed to translate the vehicle identification database into partial routes that consist of sequential equipment locations that correctly identify a vehicle plate. When a vehicle first appears in the database, the corresponding record is regarded as the starting point of the partial route. For example, the following records all belong to this trip until one record appears on a road in which the time gap between this record and the previous is large enough, or there is a movement angle under a threshold.

The process starts with each record of the identified plate. Its earlier record is also identified, and the time since the last record was calculated. Next, were calculating best-case times for this pair displacement through the in-network assignment. In the case of real-world time calculated outside a range that closely relates to actual times (with an added interval value of 6 hours), it is possible to infer that a stop occurred. Multiple intervals were tested in this dissertation, and the choice (6 hours) was selected based on which interval best matched with the surveys and expert suggestions.

An additional method applied was generating trip breaks once the trajectory of the trip shows a variation of less than  $30^\circ$  in a given direction. In this way, the outward journey is discriminated against the return journey. The threshold was selected based on an analysis of the network. Values over  $30^\circ$  would frequently characterize breaks on common paths consisting of multiple secondary accessways or mountainous roads. Figure 7a illustrates a path with turns over 30 degrees. Figure 9b illustrates a path with a turn with an angle under 30 degrees. Lastly, trips containing sub-segments with speeds over 120 kph are removed, with that being over the highest regulated allowed speed in the state. This filtering also removes trips with vehicles being wrongly identified through OCR (LPR, radars).

Figure 7: Visual simplified interpretation of path angles



Source: Author

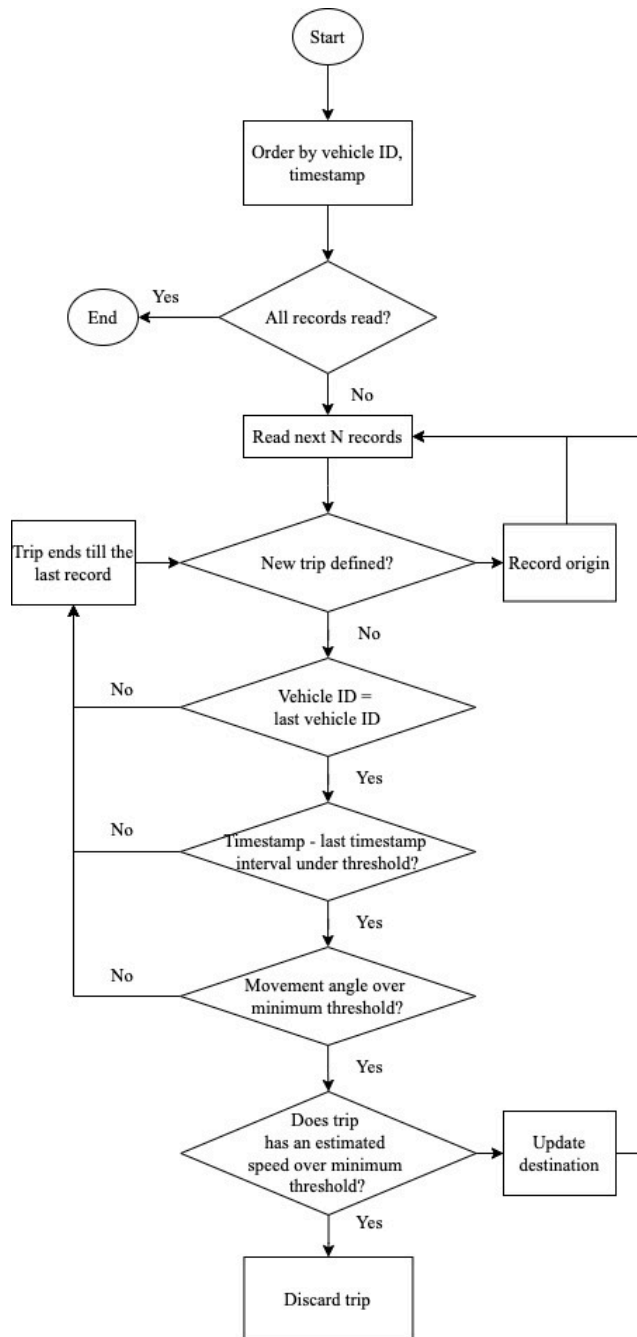
This procedure creates splits in the vehicle identification sequence and determines each split as a trip destination; sequential records are the start of new trips.

Table 3 Sample vehicle equipment sequence with a defined split based on time interval threshold

Row	Vehicle ID	Timestamp	Equipment ID	Sequence	Split	Trip
1	231523	5/10/2019 15:59	252	1		1
2	231523	5/10/2019 17:05	619	2		1
3	231523	5/10/2019 17:13	797	3		1
4	231523	5/10/2019 17:57	274	4		1
5	231523	6/10/2019 09:36	136	5	Y	2
6	231523	6/10/2019 10:31	43	6		2
7	231523	6/10/2019 10:35	451	7		2
8	231523	6/10/2019 11:07	347	8		2

In the example presented in Table 3, vehicle 231523 goes through the equipment sequence presented in the “Equipment id” column. A significant amount of time has passed at the event in row five, marked as Y in the Table 3. With a split defined, the procedure breaks the sequence into two trips, the first beginning at id 252 and ending at 274 and a second trip, beginning at 136 and ending at 347. Figure 8 shows the general algorithm structure.

Figure 8: A flow chart explaining the origin-destination extraction from AVI data



Source: Author

First, all data points are ordered by vehicle ID and timestamp. Then, the algorithm compares two sequential records. If they have the same ID, its timestamp difference is less than the threshold (best-case network time added with 6 hours), movement speed is below a maximum speed threshold (e.g., 120 kph), movement is below the angle threshold, then these records belong to the same trip, and the destination is updated to the last record. Otherwise, the former record is the destination of this trip, and the next record is the origin of the next trip. This way, the origin and

destination of each partial route and their corresponding location and time are extracted. Zhang (2019), published after this method was created, had a similar approach, with a few different assumptions.

After the steps in this section, the method builds a database of equipment sequences (partial routes), grouped and summarized by vehicle category (bus, trucks, cars). By selecting its start and end-points, origin demand matrices are extrapolated. While these might seem like enough for a trip assignment step, start and end equipment are not directly correlated to start and end zone. Selecting the closest zones from each start and endpoint could give substantial errors to the zone matrices representing actual origins and destinations of trips. This circumstance creates the need for an additional step for determining the probable origin and destination zones with a more robust method, outlined in section 4.2.

#### **4.2. O/D Matrix construction (Step 2)**

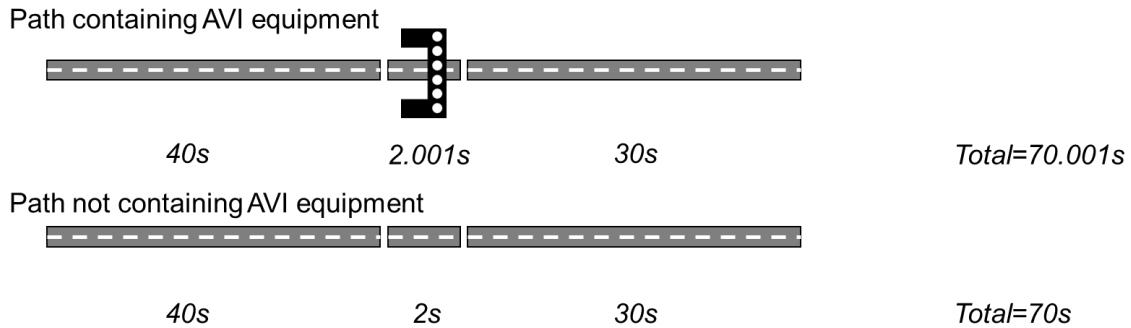
As previously outlined in the earlier section, although the partial route identification can translate vehicle path data, it is limited to an assignment of actual O/D demand on the road network present at the time of data collection. Therefore, any O/D estimation based solely on it would not be appropriate for scenario-based studies, where different modes of transportation or road network changes are proposed.

Therefore, it is necessary to establish a way to select and ponder whether some generic zone should be or not responsible for a proportion of the total vehicle flow on each partial route. In this scenario, each zone has a weight attributed to itself. This paper chose a premise of population and employment having equal contribution. Furthermore, a gravitational model was built using these weights and a variable of cost, attempting to fulfill the need for a probabilistic distribution of traffic flow.

An algorithm was created that selects a set of likely origin and destination zones for each partial route generated. First, some essential points were established: the method depends on the minimum cost path assignment and the high reliability of ETC systems. In other words, it is postulated that if an ETC capture system did not detect a vehicle, it did not pass through that section. Then, they added a decimal fraction to a whole

network cost value allowed to different paths with or without ETC sections by their added final cost (Figure 9).

Figure 9: Visual representation of decimal costs on ETC equipped network links



Source: Author

Identifying paths containing decimal components can detect trips that travel through ETC equipment. For a set of candidate origin or destination zones for each partial trip, any trips with a decimal portion would have their zone removed from the candidate zones. Table 4 gives an example of a trip with three candidate zones (A, B, and C). By analyzing zone B path cost, it is determined that it should have to cross an ETC for the vehicle to reach zone B. Therefore, the path and corresponding zone (B) are removed, and the information is not recorded in the database. The set of candidate zones for the vehicle is A and C.

Table 4 Method to filter out paths that go through ETC detectors

	Path to zone A	Path to zone B	Path to zone C
link 1	2		
link 2		3	3
link 3	4		4
link 4 (AVI)		0.01	
link 5	7	7	7
link 6		3	3
path total cost	13	13.01*	17

Source: Author

Another filtering process is eliminating origin and destination pairs with lower-cost path alternatives, thus eliminating improbable route choices.



Each unique trip is associated with an origin zone, then to the first AVI detector, then the last AVI detector, and finally its destination zone. This way, for each pair of first and last sensors the vehicle passed through, there is a set of possible origin and destination zones, based on the weights of each zone and the cost associated with the pair. The portion of the total combined trip count for the partial route segment and each of its origin and destination associated pairs are distributed following a classical gravity model of travel distribution. Following the model, force of gravity is more significant for a large object and small for a large distance. The classical model is as follows – Equation (1):

$$G = g \frac{Mm}{r^2} \quad (1)$$

Where:

G: the force of gravity between two objects

g: gravitational constant;

M,m: object mass;

r: distance between two objects

Based on the notion of gravity, we can assume that a higher share of vehicles from a specific partial route comes from zones with higher attraction. Subsequently, the gravity flow model is established and indicated in Equation (2). Finally, the object mass is replaced with a sum of population and employment, each having equal contribution, and the radius is replaced with a time-based generalized trip cost.

$$Volume\ Split_{i,j}^n = \frac{\frac{(P_i + E_i) \times (P_j + E_j)}{c_{ij}^2}}{\sum_{a \in z_o^n, b \in z_d^n} \frac{(P_a + E_a) \times (P_b + E_b)}{c_{ab}^2}} \quad (2)$$

Where:

$z_o$ : Selected origin zones for partial route  $n$ ,

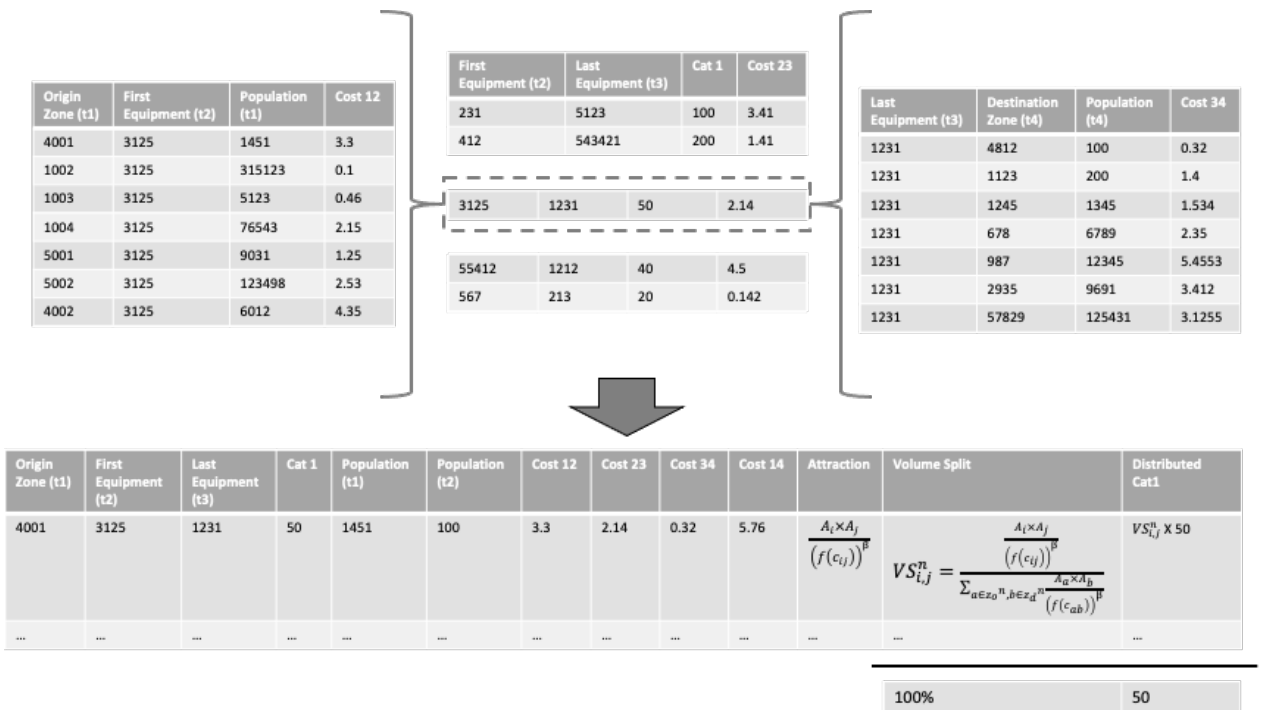
$z_d$ : Selected destination zones for partial route  $n$ ,

$c_{ij}$ : Generalized trip cost,

$P_i$ : Population for zone  $i$ ,  
 $E_i$ : Employment for zone  $i$ .

The algorithm splits the partial trip dataset into chunks, joins tables containing all likely start and end zones (out of all 1.054), and then redistributes trip volumes with the calculated volume splits or probabilities. The complete process is exemplified in Figure 10. For a partial trip defined as having the first trip detection being in equipment 3125 ( $t_2$ ) and the last being 1231 ( $t_3$ ), we can determine a set of probable starting zones ( $t_1$ ), each having its attributed costs ( $c_{12}$ ) between the origin and the first partial route equipment, as well as a set of probable end zones ( $t_4$ ), each with its attributed costs ( $c_{34}$ ) between the last partial trip equipment and the destination zone. The cost of the whole trip ( $c_{14}$ ) is calculated by adding each sub-segment ( $c_{12} + c_{23} + c_{34}$ ).

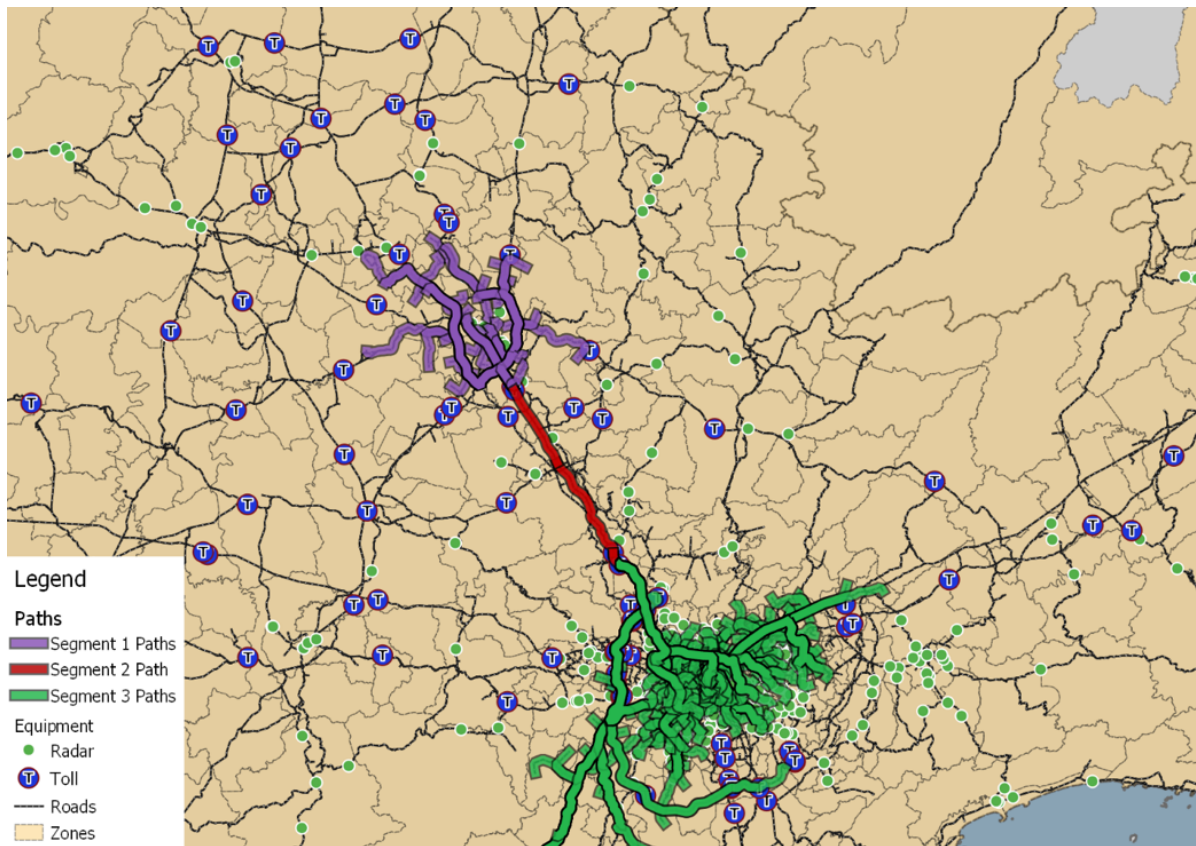
Figure 10: Visualization of the matrix estimation method



Source: Author

In practice, Figure 11 shows this step of the project: in purple are the routes that begin in a range of origin zones and have a high probability of passing, initially, by a specific toll; in red is the path from the first ETC equipment to the last one; in green are the most likely paths from that last equipment to a range of specific destination zones.

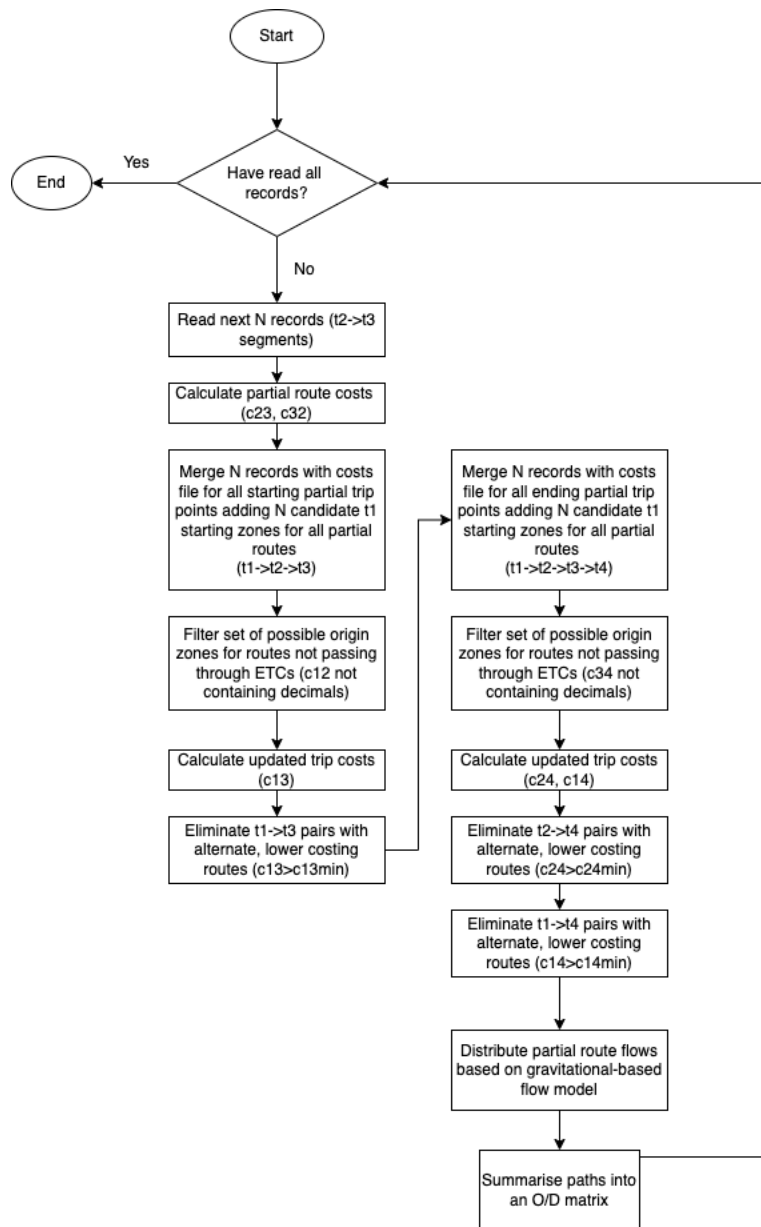
Figure 11: Result of a set of possible origin and destination zones for the selected partial route segment



Source: Author

Figure 12 shows the algorithm of OD estimation and flow distribution, and section 4.2.1 details the commented algorithm

Figure 12: A flow chart explaining the algorithm for extrapolating network OD and distributing trip flows under a gravitational model



Source: Author

### 4.3. Matrix calibration (Step 3)

The final activity group uses the *T-Flow fuzzy* algorithm present in the commercial software VISUM from PTV. The algorithm uses vehicle counts as input, correcting origin and destination pair total amount of trips to represent these volumes correctly. The steps in the algorithm are as follows.

### 4.3.1. Data Input

The first step is to observe the travel demand for origin-demand pairs, which describes the trip demand pattern of a previous state. Multiple origin-destination pairs could have a contributing share of trips for each traffic count. The counted volumes account for the sum of all O-D pairs traveling on this specific link. The initial O-D matrix and the observed link counts are the input data for the algorithm. The first step of iterations is based on existing information of link flows and O-D demand.

The notations of variables used in this step are presented as follows:

$\varepsilon$ : Maximum allowed change (percentage of matrix totals) of O-D estimated at each consecutive successful iteration.

$n$ : Iteration counter.

$T_{ij}^{n=0}$ : Number of trips from origin  $i$  to destination  $j$  in the starting O-D matrix.

$P_{ij}^n$ : Link share from the calculated number of trips from origin  $i$  to destination  $j$  at iteration  $n$ .

### 4.3.2. OD matrix assignment

At every iteration ( $n$ ), the objective matrix related to that iteration ( $T_{ij}^n$ ) is assigned to the transportation network by the User Equilibrium (UE) traffic assignment model, first proposed by Wardrop (Sheffi, 1985). With the calculated flows on network links by implementing the UE model, route choice shares ( $P_{ij}^n$ ) are then estimated. In transportation modeling, user equilibrium describes a route choice assumption proposed by Wardrop: “The journey times on all the routes actually used are equal and less than those which would be experienced by a single vehicle on any unused routes”, also known as Wardrop’s first principle.

Next, the flows for each O-D pair are loaded onto the network based on the travel time (or impedance) of the alternative paths that could carry this traffic. The algorithm says that flows on links are in equilibrium when no user can improve his travel time by unilaterally shifting to another route choice.

### 4.3.3. OD matrix calibration

The mathematical formulation for the O-D estimation problem considered in this research is shown below, Equations (3), (4) and (5).

$$Max - \left( \sum_{ij \in OD} T_{ij}^n \ln \left( \frac{T_{ij}^n}{t_{ij}} \right) - T_{ij}^n \right) \quad (3)$$

Restricted by:

$$\sum_{ij \in OD} T_{ij}^n \times P_{ij}^n = \tilde{v}_a; \forall a \quad (4)$$

$$T_{ij}^n > 0 \quad (5)$$

Where:

$\varepsilon$ : Maximum allowed change of O-D estimated at each consecutive successful iteration.

$n$ : Iteration counter.

$T_{ij}^n$ : Number of trips from origin  $i$  to destination  $j$  in the O-D matrix from iteration  $n$ .

$t_{ij}$ : Number of trips from origin  $i$  to destination  $j$  in the initial target O-D matrix.

$\tilde{v}_a$ : Observed traffic count on the link with variable bandwidth.

$P_{ij,a}^n$ : Route choice proportions for link  $a$  at iteration  $n$ .

### 4.3.4. Convergence criteria

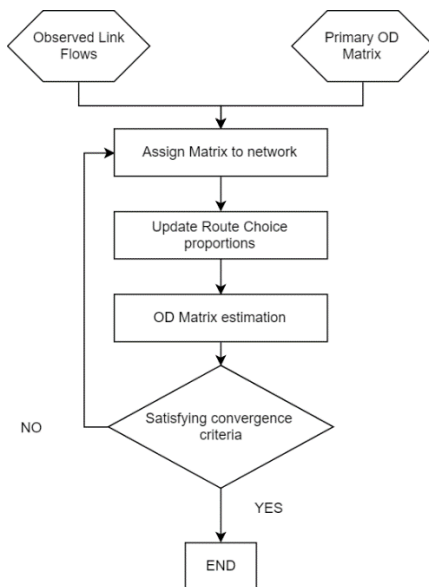
The convergence criteria are based on the acceptable difference between the estimated matrix and the previous estimation step matrix (Step 1 and Step 2). If not met, the technique (T-Flow fuzzy algorithm) continues with redoing the previous steps, though, with the basic fact that the route choice proportions are updated by assigning the newly estimated matrix and not the initial target matrix.

$$E = \sum_{ij} (T_{ij}^n - T_{ij}^{n-1})^2 \quad (6)$$

If  $E < \varepsilon$  stop, otherwise  $T_{ij}^n = \frac{T_{ij}^n + T_{ij}^{n-1}}{2}$  Moreover, go to data input.

This step is exemplified in Figure 13 with a diagram of the algorithm's iterative nature.

Figure 13: T-Flow fuzzy diagram



Source: Author

Yousefikia et al. proposed a modification in the T-Flow fuzzy algorithm in which the route choice proportions are updated successively at each iteration, allowing for a more precise estimation of the OD matrix [46]. However, the modified T-Flow fuzzy was not applied in this dissertation.

#### 4.4. Chapter final remarks

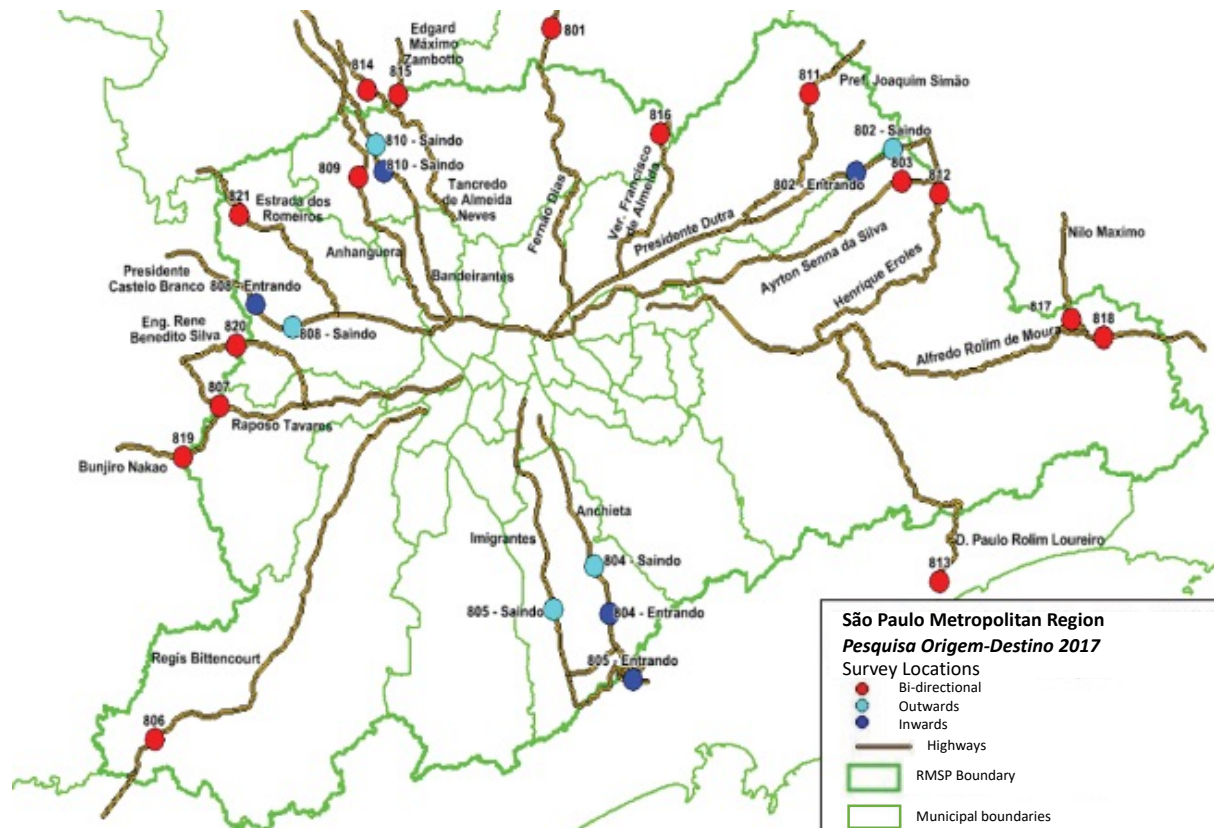
This chapter presents the proposed method to fulfill the objectives outlined in this dissertation. It brings its steps, structure, organization, how it plans to employ the different datasets, and how they are treated. A more robust and well-grounded technical approach to estimating attraction vectors could yield great improvements in the OD distribution analysis, consisting a subject for further study.

## 5. A Method for pre-Validation

Following the development and application of the estimation algorithm, in conjunction with experts and academic advisors, a method for validating the result of extrapolating and distributing partial routes was proposed.

The method compared aggregated zone demand with data from the *Pesquisa Origem-Destino 2017*, carried out by the *Companhia do Metropolitano de São Paulo*. For each survey location, with an accompanied AVI-equipped toll plaza, we compared model and survey captured trips. Figure 14 displays the survey locations from *Pesquisa Origem-Destino 2017*

Figure 14: Survey locations, *Pesquisa Origem-Destino 2017*



Source: Pesquisa Origem-Destino 2017

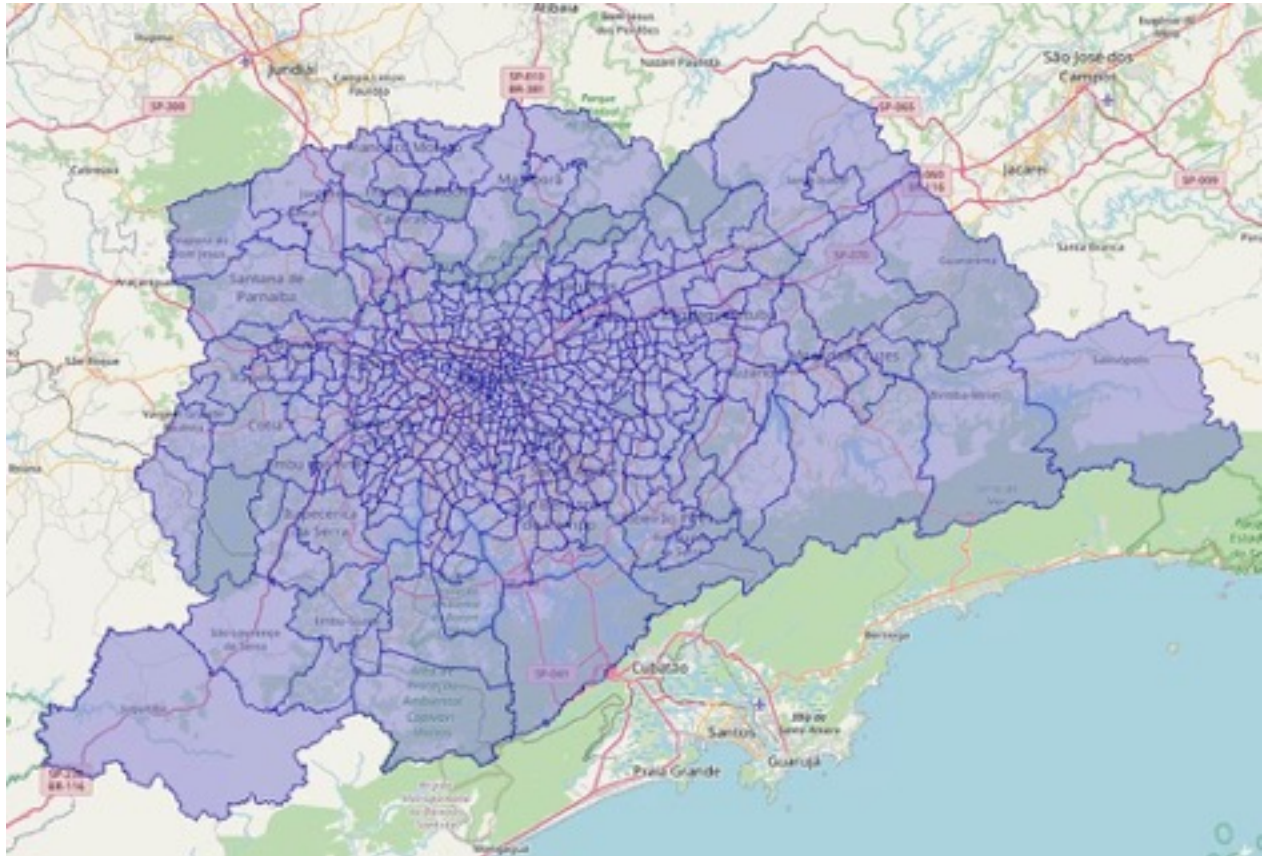
### 5.1. Zone aggregation

The essential primary step was to establish a correlation and aggregation of zones between both sources. Unfortunately, the 517 zones defined in the survey (Figure 15) did not match the 1058 zones defined in this dissertation (Figure 16). Nevertheless,



the significantly high level of detail was well-suited to establish an aggregated comparison of total origin and destination demand analysis.

Figure 15: Zones, *Pesquisa Origem-Destino 2017*



Source: (Costa, Breno, 2021)

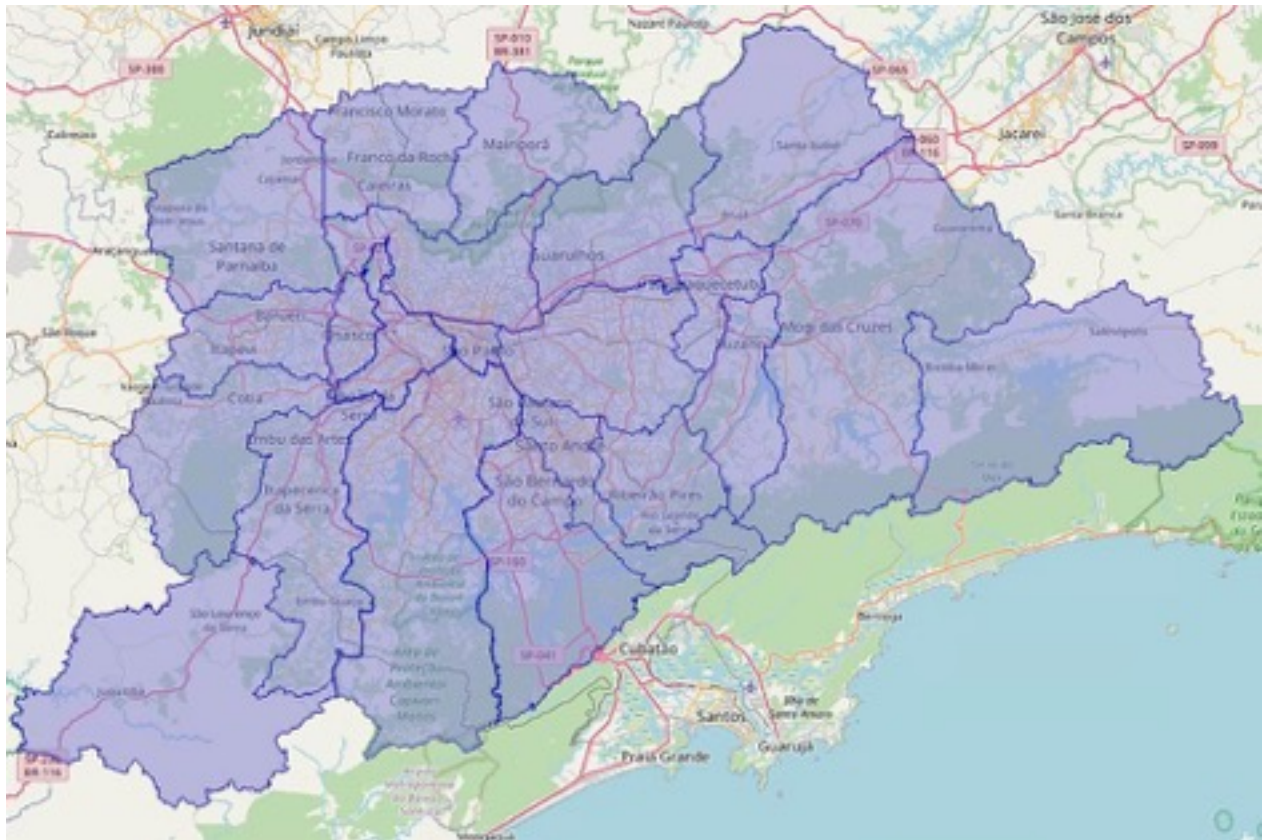
Figure 16: Zones, Current Dissertation



Source: (Costa, Breno, 2021)

The aggregation method followed transportation planning best practices, focusing on zones higher in demand and inside the metropolitan region of the State of São Paulo. Zones within the metropolitan region had municipalities with closely tied transportation networks merged, except for the state capital, São Paulo, that due to its size was aggregated between its five zones (north, south, east, center, west). Figure 17 shows the result of this aggregation step, with 22 zones defined.

Figure 17: SP metropolitan zone aggregation



Source: (Costa, Breno, 2021)

With the focus of the survey being within the metropolitan region and the lack of detailed data from outside this boundary, we decided to aggregate zones to their respective microregions<sup>4</sup>. One exception was the *Baixada Santista* region, with its closely tied transportation networks between its municipalities. However, the fact that they are both located within the island of São Vicente required a different approach, detailed in Figure 18.

---

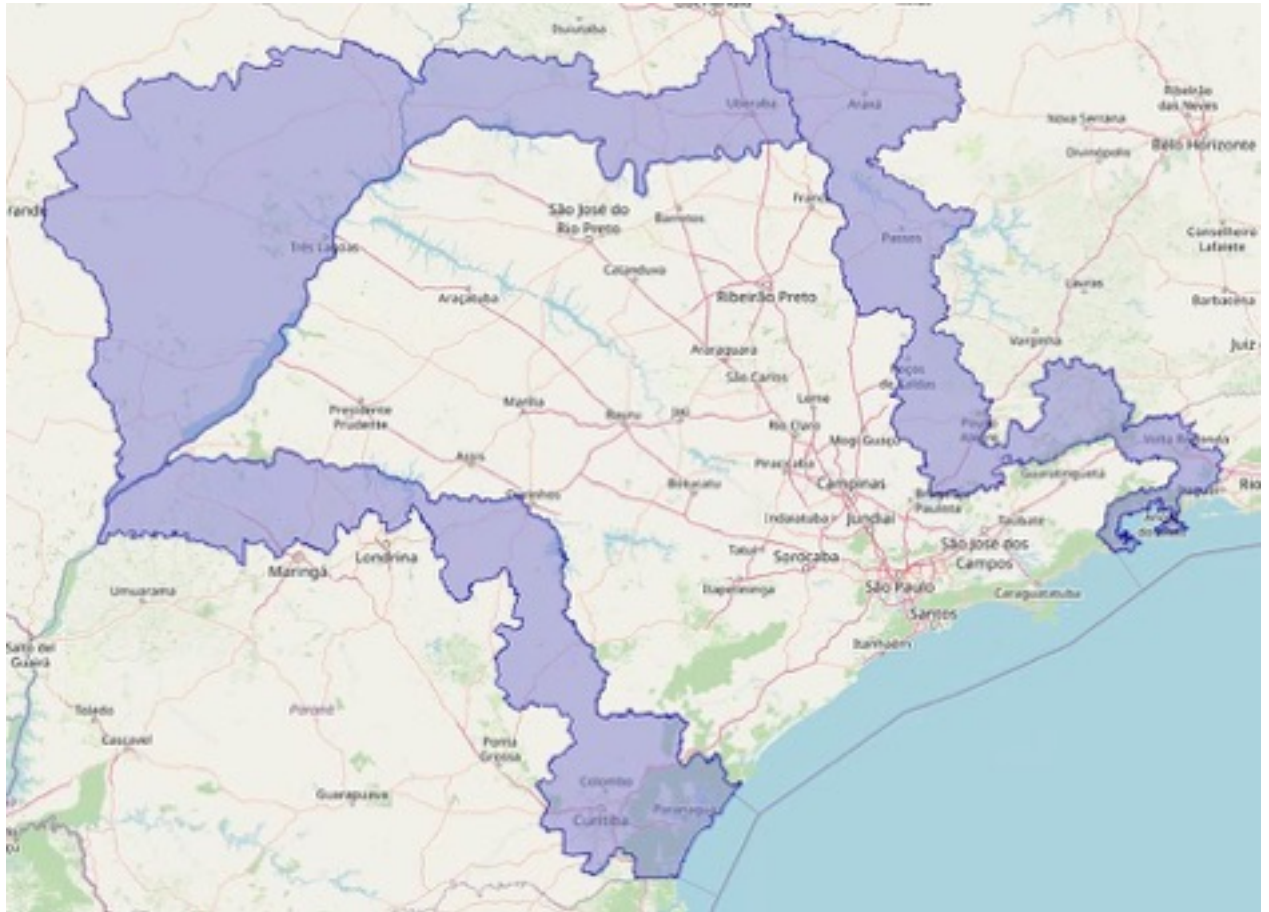
<sup>4</sup> [https://pt.wikipedia.org/wiki/Mesorregi%C3%B5es\\_e\\_microrregi%C3%B5es\\_do\\_Brasil](https://pt.wikipedia.org/wiki/Mesorregi%C3%B5es_e_microrregi%C3%B5es_do_Brasil)

Figure 18: *Baixada Santista* zone aggregation

Source: (Costa, Breno, 2021)

To represent the were long-distance trips in the remainder of the country, we adopted the criteria based on the direction of trips outside the boundaries of the dissertation. There are a few main axes of transportation coming from and to the State of São Paulo, and these can be represented by three zones, represented in Figure 19. These are the Southeast-Northeast zone, which includes roads that connect the state to the remainder of the Southeast region and the Northeast region. The center-west zone includes connections between the study area and the Center-west and North regions. Moreover, the final third is the South zone, interlinking the State of São Paulo with the South of Brazil.

Figure 19: Border aggregate regions



Source: (Costa, Breno, 2021)

In conclusion, Figure 20 shows the complete aggregation process, resulting in 82 zones.

Figure 20: Final aggregated zones



Source: (Costa, Breno, 2021)

## 5.2. Assignment and R-squared analysis of model and survey data

Following the aggregation of zones, we validated the partial results coming from the matrix estimation step with data from the *Pesquisa OD 2017*. In this section we aim to compare the fraction from the total flows coming from the surveys, to the results of the zone extrapolation.

A characteristic of the main highways in the State of SP is that they are radial concerning the City of SP. For this reason, the analysis began separately for each highway – *Rodovia dos Imigrantes*, *Rodovia Anchieta*, *Rodovia dos Bandeirantes*, *Rodovia Presidente Castelo Branco* e *Rodovia Anhanguera*. The choice of these locations came since these highways represent the principal axes of transportation in the State of São Paulo, by each connecting with their four closely tied regions and transportation lanes to every other region in the country. Figure 21 explains this quadrilateral distribution in the region.

Figure 21: State of São Paulo Metropolitan Quadrilateral



Source: Author

An analysis of the R-square was performed between the Metro Survey data and the distribution proposed by the model. The initial procedure was aggregating trips from the estimated zone matrix (with 1,054 zones) into the proposed 82 zones by origin and comparing the contributions of each zone with an aggregation of the Metro Survey results (517 zones) into the same 82 zones. The analysis began separately for each highway – Rodovia dos Imigrantes, Rodovia Anchieta, Rodovia dos Bandeirantes, Rodovia Presidente Castelo Branco e Rodovia Anhanguera. The choice of these locations came since these highways represent the principal axes of transportation in the State of São Paulo, by connecting with their four closely tied regions and transportation lanes to every other region in the country.

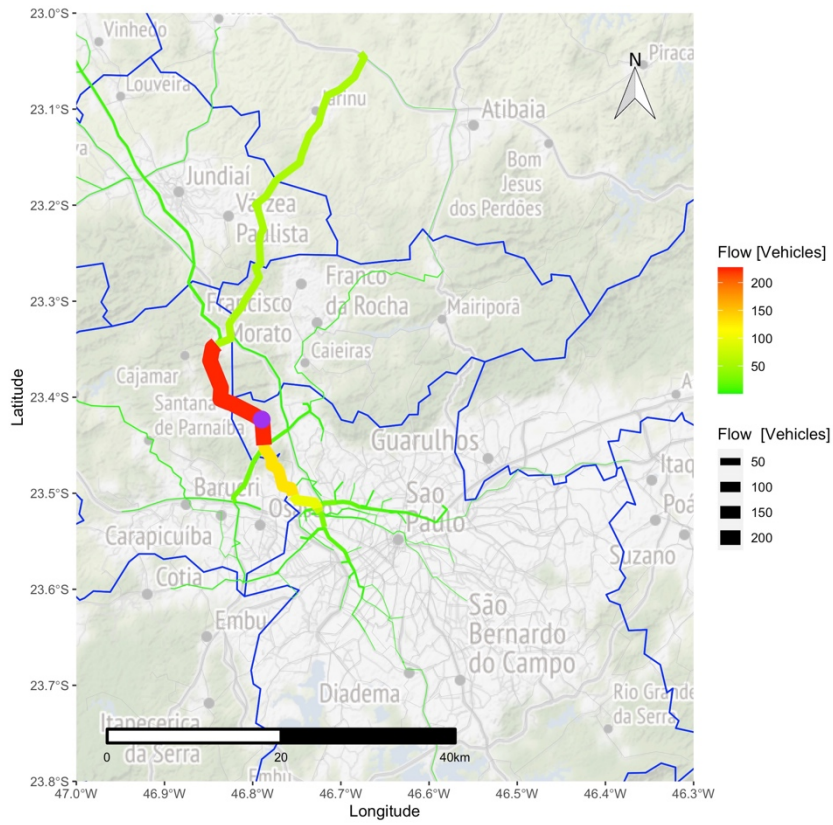
In the following sections, we discuss the results from the R-squared analysis for each of these axes and then an aggregated analysis of all data points.

### ***Sistema Anhanguera-Bandeirantes, São Paulo - Campinas***

The aggregate of *Rodovia dos Bandeirantes* with *Rodovia Anhanguera* is defined as the *Anhanguera-Bandeirantes* System, as of today being administered by AutoBan. Considered as one of the best maintained in the country and holding significant commercial relevancy, as together with the *Rodoanel Mário Covas* and the *Rodovia Anchieta*, acts as the connection between two of the most important import and export centers in the country: the Viracopos International Airport and the Port of Santos. One other reason it holds so much significance is that it connects the two wealthiest metropolitan regions in the country: São Paulo and Campinas. Figure 22 to Figure 25 show the result of the assignment prior and after the trip endpoint estimation algorithm for *Rodovia dos Bandeirantes* and *Rodovia Anhanguera*, respectively. These figures show the distribution effect and have guided the interpretation of data and the functionality of the procedure. Since this step only applied all or nothing assignment, as a tool of visual interpretation of general algorithm function, we can expect to see transference of trips from *Bandeirantes* to *Anhanguera* (or vice-versa) due to the close competition between both options.

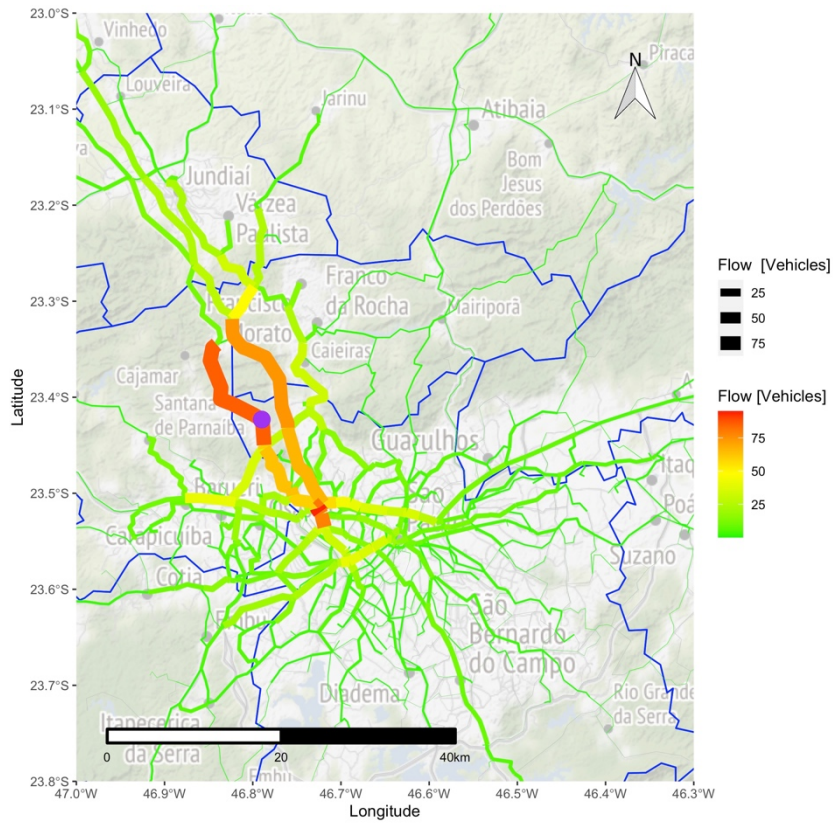


Figure 22: Rodovia dos Bandeirantes, trip distribution, partial routes



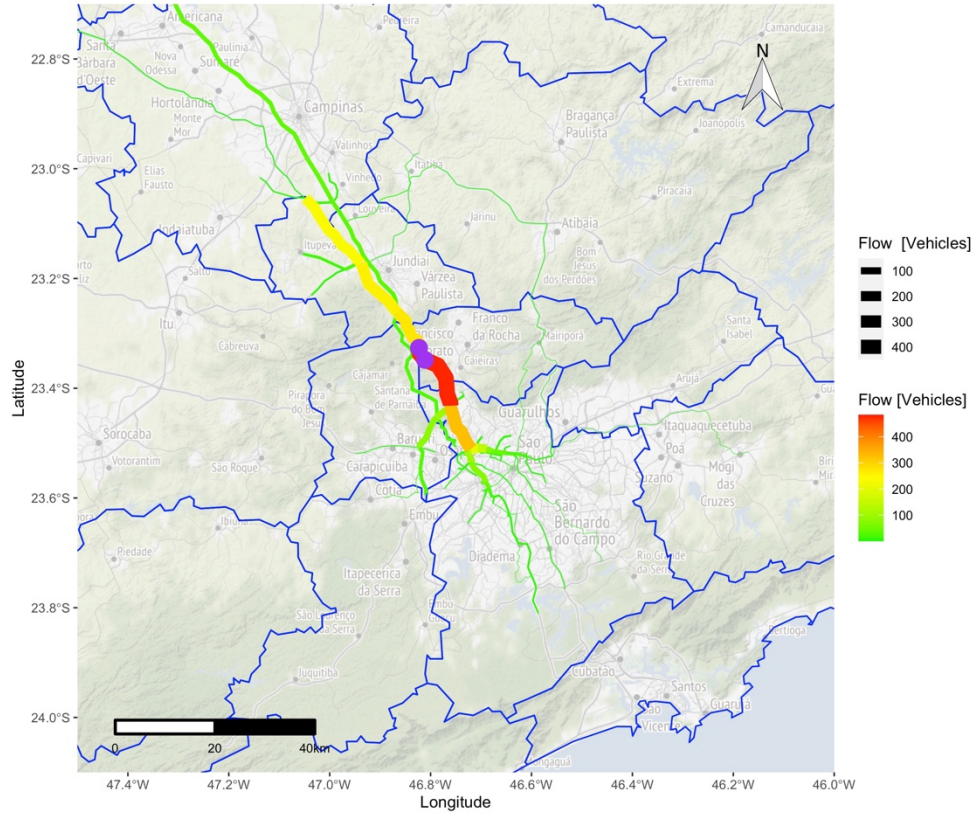
Source: Author

Figure 23: Rodovia dos Bandeirantes, trip distribution, zone extrapolation



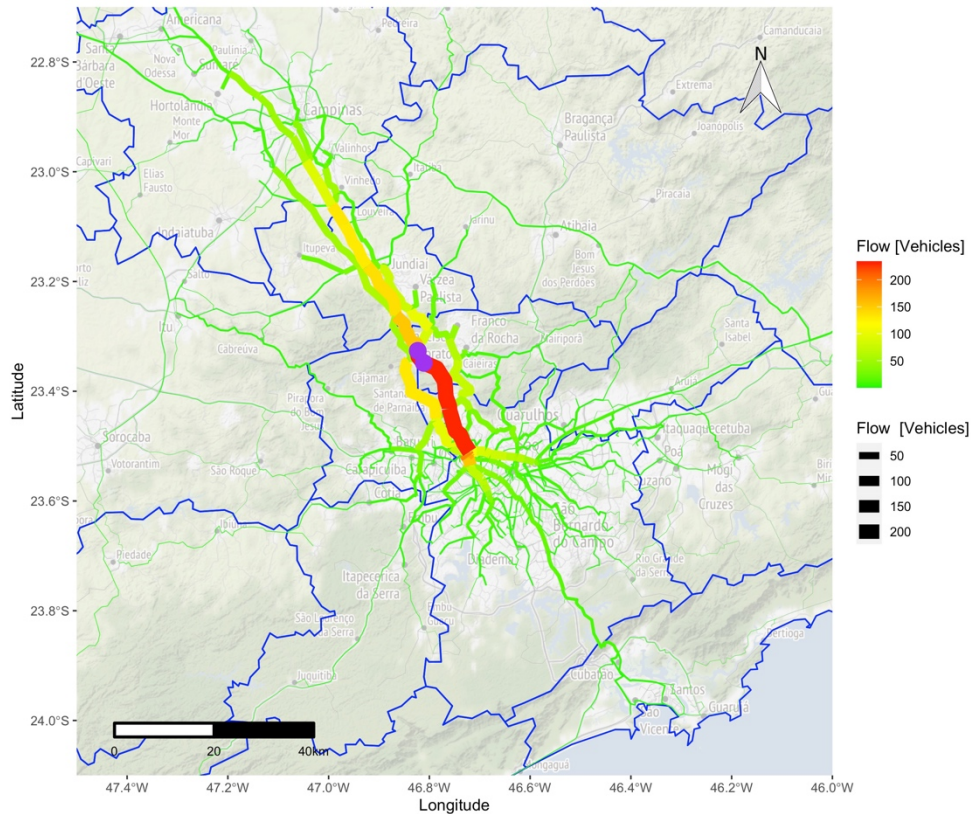
Source: Author

Figure 24: Rodovia Anhanguera, trip distribution, partial routes



Source: Author

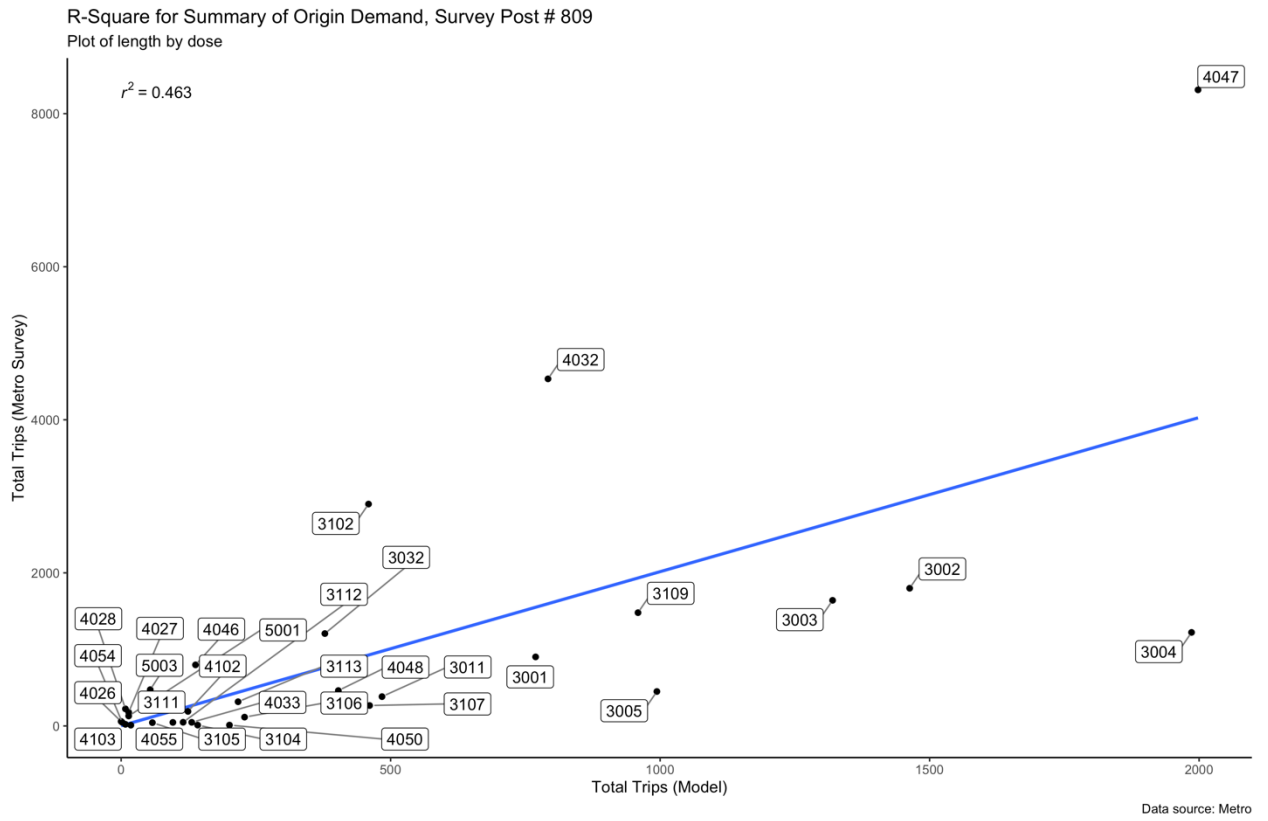
Figure 25: Rodovia Anhanguera, trip distribution, zone extrapolation



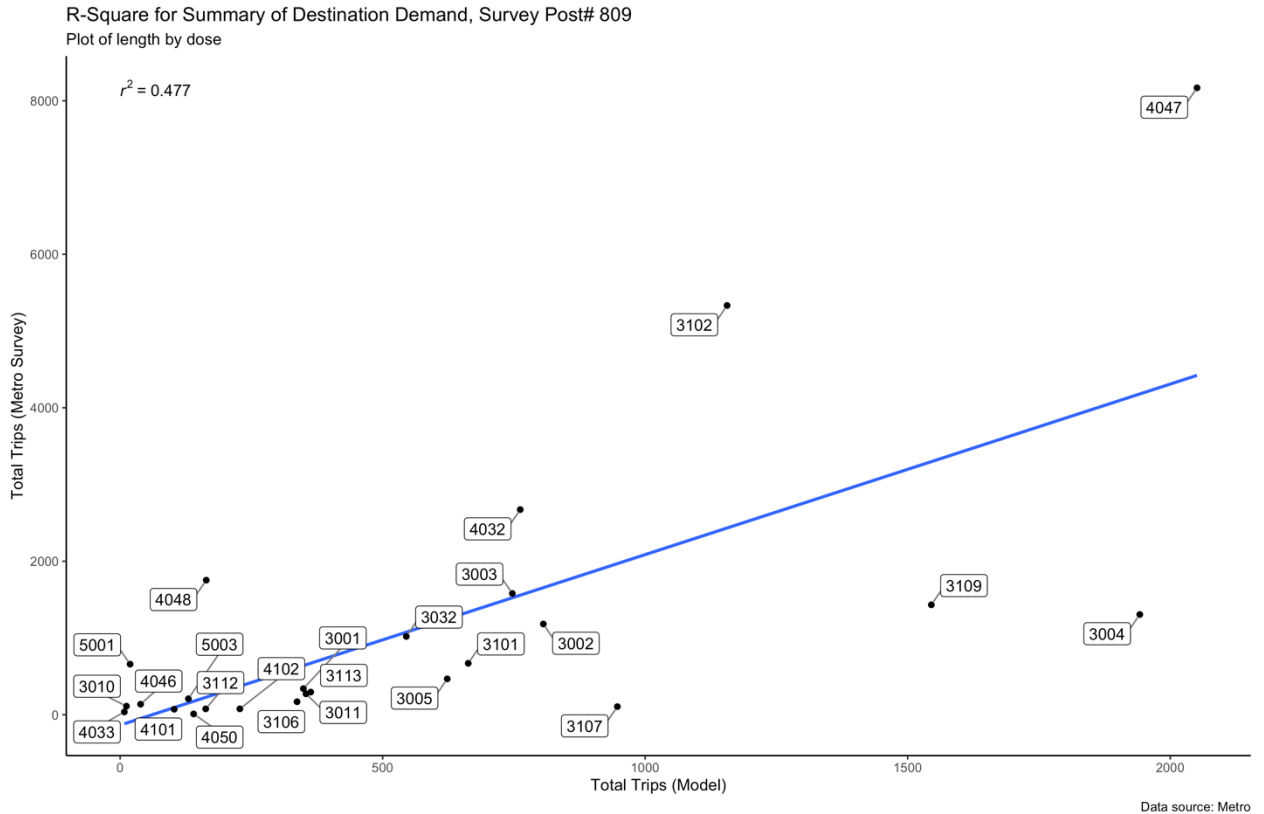
Source: Author

Afterward, we could plot the model outputs in a sequence of images from the summarized perspective or origin and another from destination demand. For example, Figure 26 and Figure 27 shows the R-squared analysis and result for OD survey outpost #809, Rodovia dos Bandeirantes. R-squared achieved was 0.463 and 0.477, respectively.

Figure 26: R-squared analysis of origin demand in outpost 809 (Anhanguera)



Source: Author

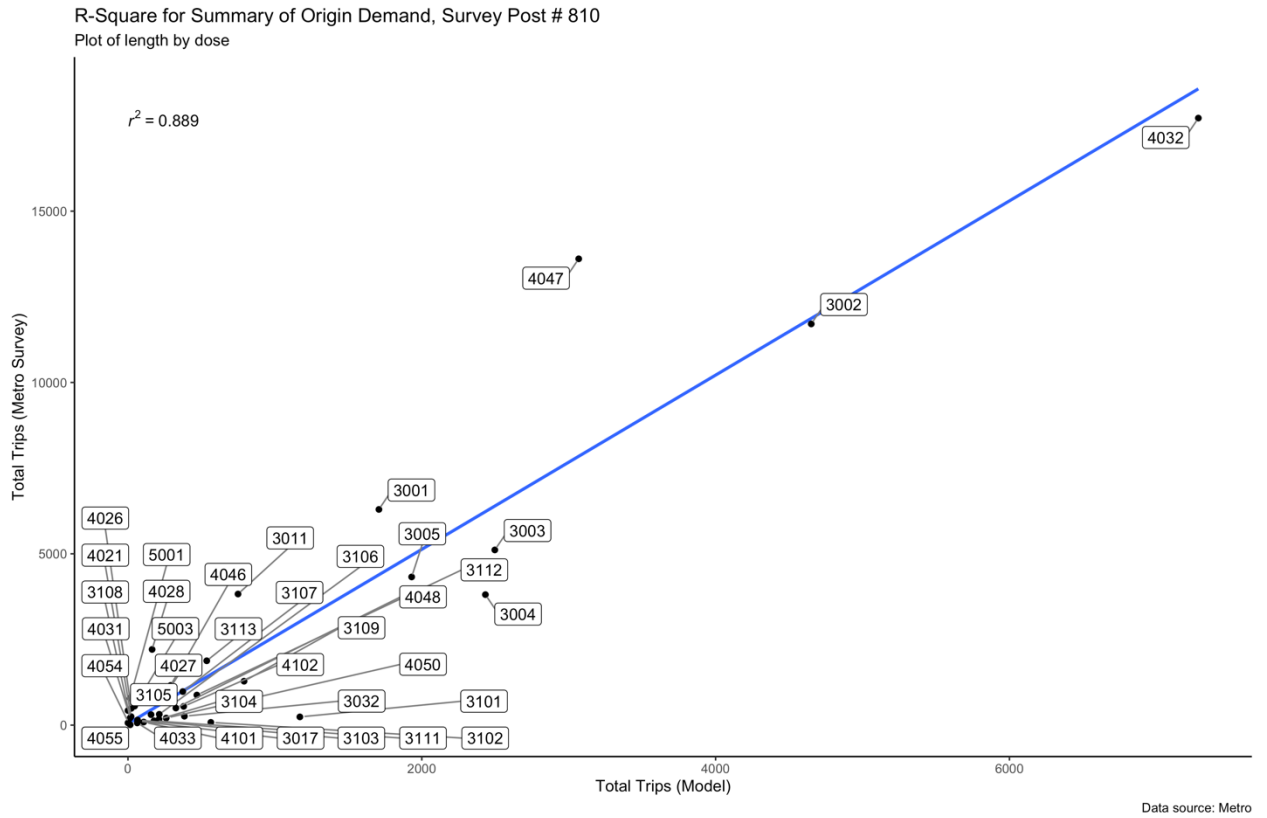
Figure 27: R-squared analysis of destination demand in outpost 809 (*Anhanguera*)

Source: Author

The results for *Rodovia Anhanguera* were under this dissertation's expectations and could not be explained by a lack of sufficient data. Instead, the interpretation of the low accuracy came from the lack of network and regional detail on the smaller cities located between the Sao Paulo and Campinas axis, with a specific issue with the municipality of Cajamar and its attraction vectors exercising an over-represented attraction of demand. The results show the limitations of modeling demand attraction based solely on population and employment, thus recommending further studies to include other demand modeling effects.

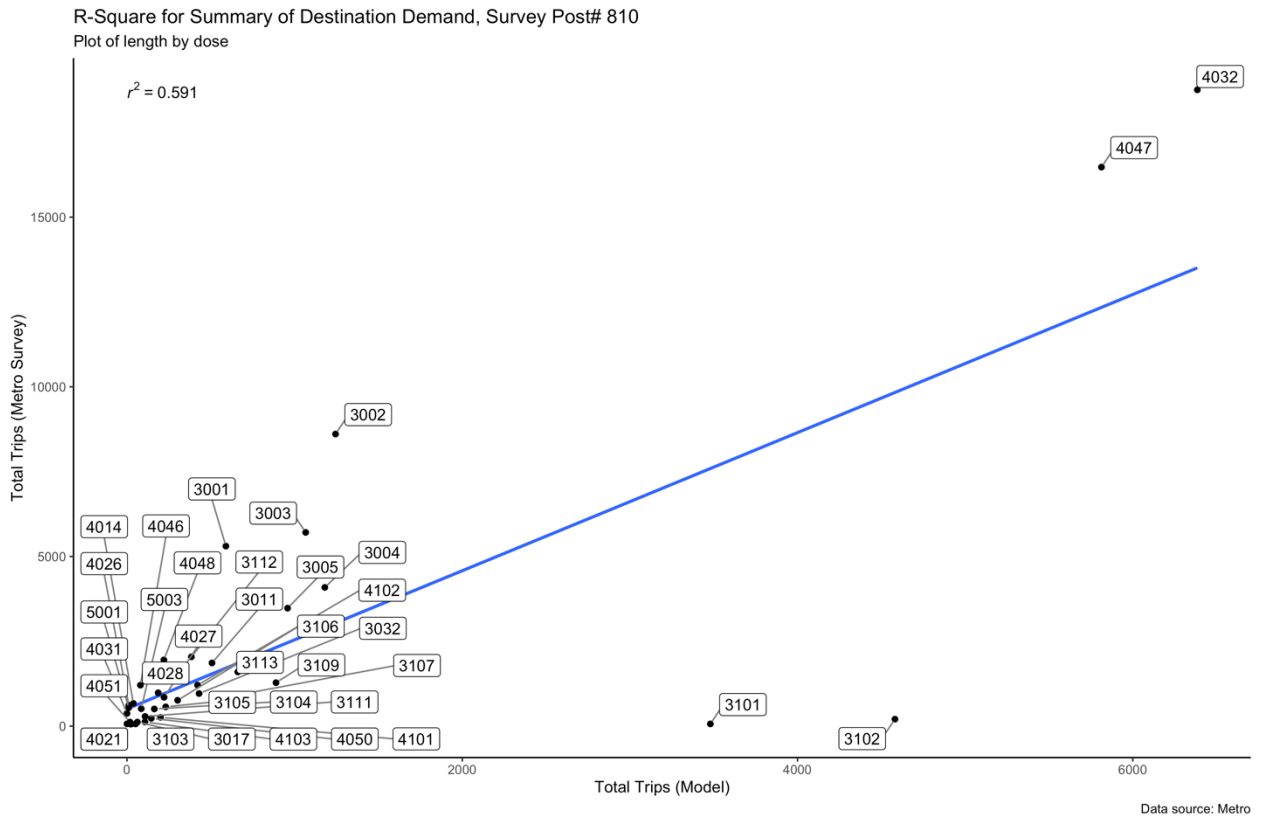
Figure 28 and Figure 29 show the R-squared analysis and result for OD survey outpost #810, *Rodovia dos Bandeirantes*. R-squared achieved was 0.889 and 0.591, respectively.

Figure 28: R-squared analysis of origin demand in outpost 810 (*Bandeirantes*)



Data source: Metro

Source: Author

Figure 29: R-squared analysis of destination demand in outpost 810 (*Bandeirantes*)

Source: Author

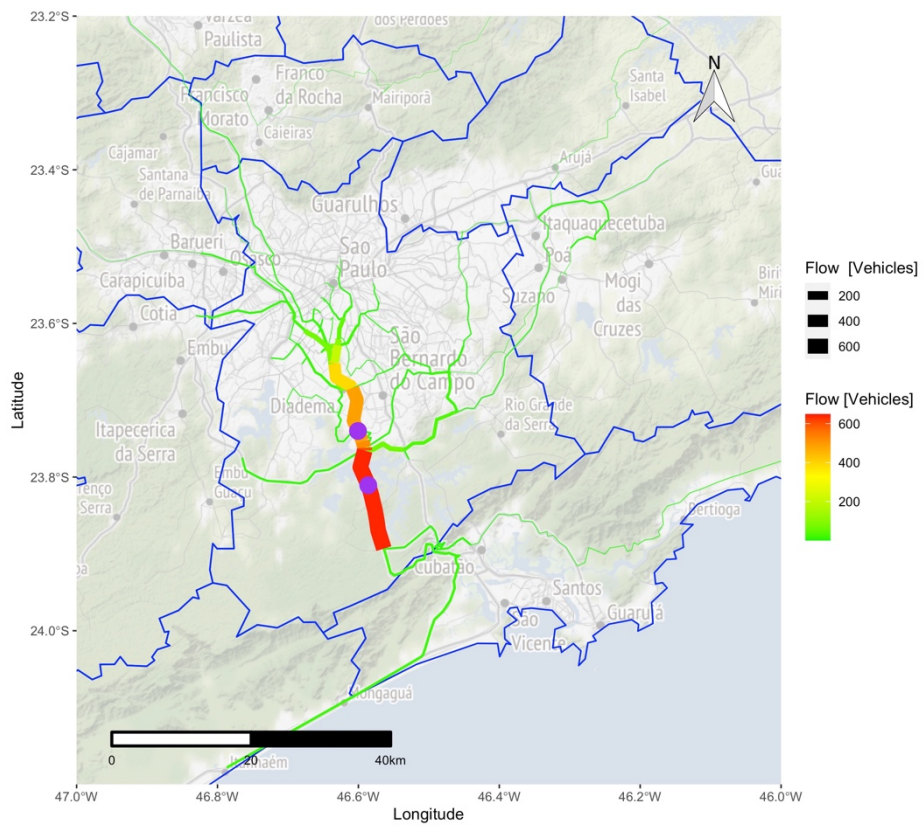
As presented in the previous plots, the result from Rodovia dos Bandeirantes shows sufficient accuracy and is more in line with this dissertation's expectations. Furthermore, the reason is consistent with the issues brought on by the municipality of Cajamar, which in the case of Bandeirantes, does not hold a connecting network element, thus discounting its harmful effect.

### ***Sistema Imigrantes-Anchieta, São Paulo - Santos***

The *Imigrantes-Anchieta* system, known as SAI, comprises the SP-160 (*Imigrantes*) and SP-150 (*Anchieta*), with both crossing the *Serra do Mar*, the plains-highland connection between São Paulo and the coastal region of the state. As previously noted, it holds significant commercial relevancy tying the totality of Brazil to the Port of Santos, its most important marine import-export port. Figure 30 to Figure 33 show the result of assignment prior and after the trip endpoint estimation algorithm for *Rodovia dos Imigrantes* and *Rodovia Anchieta*, respectively.

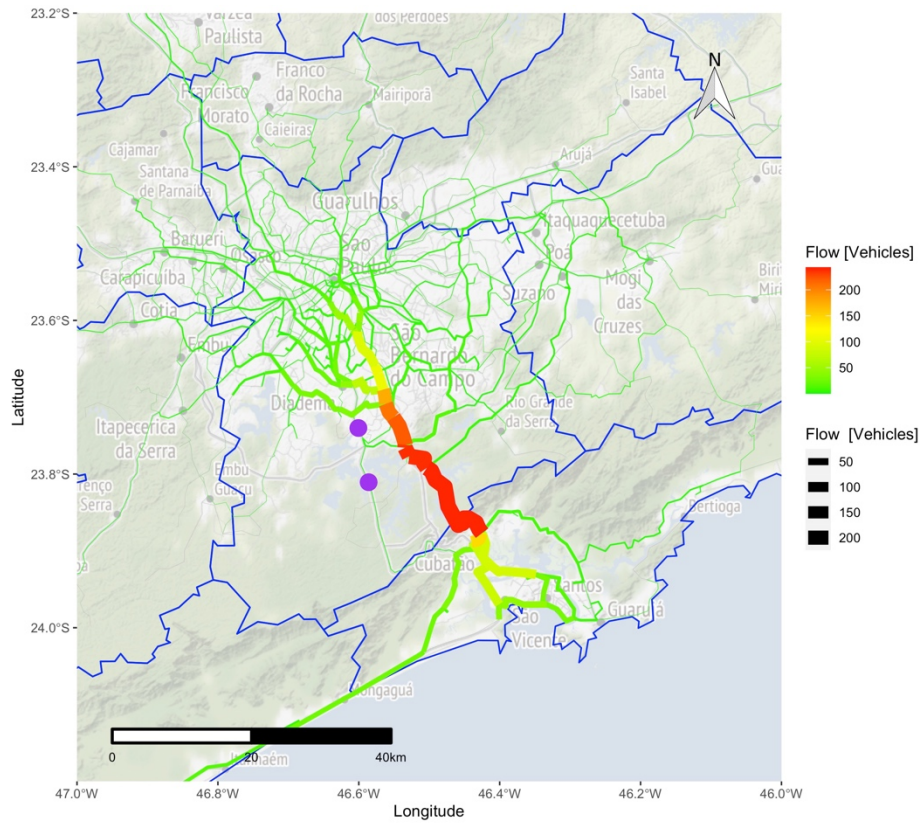


Figure 30: Rodovia dos Imigrantes, trip distribution, partial routes



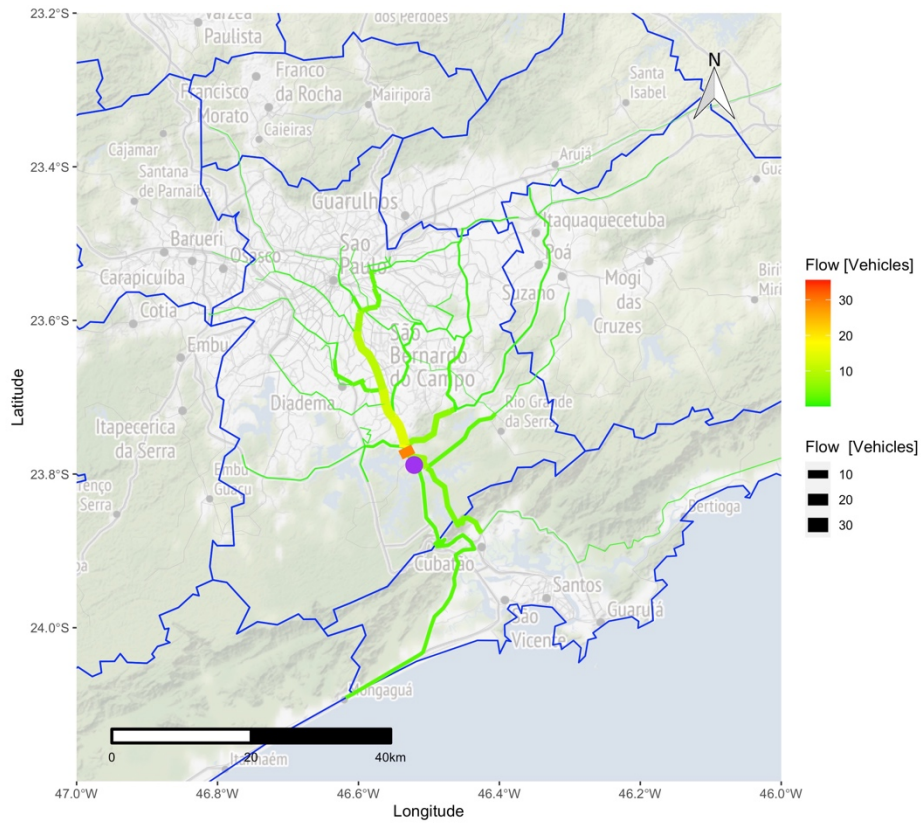
Source: Author

Figure 31: Rodovia dos Imigrantes, trip distribution, zone extrapolation

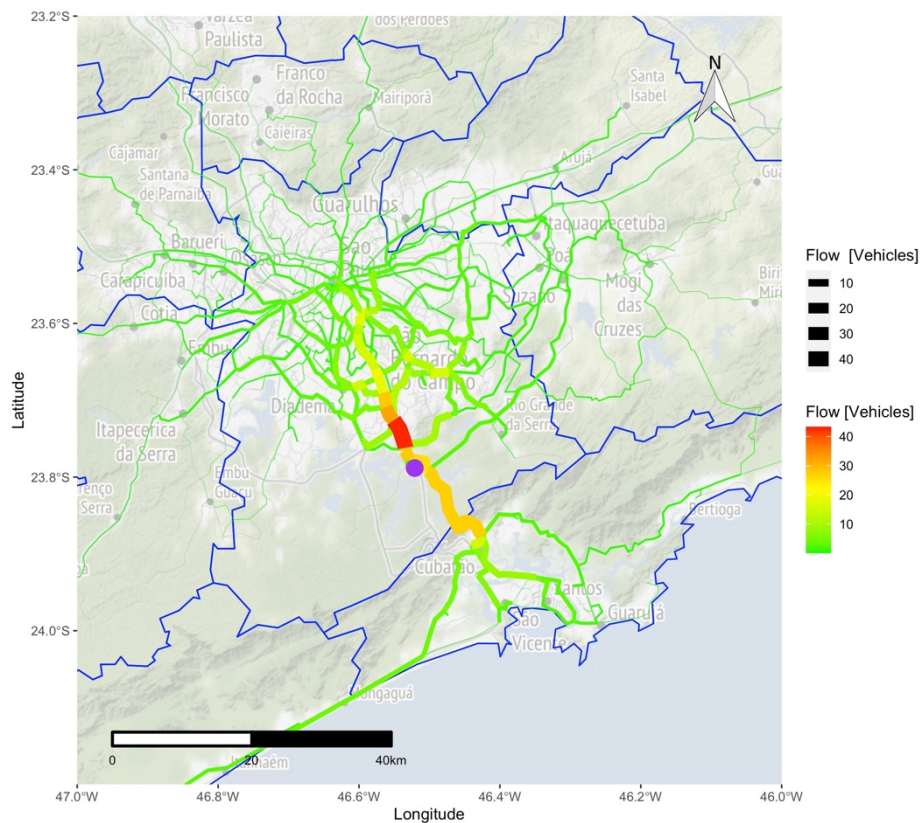


Source: Author

Figure 32: Rodovia Anchieta, trip distribution, partial routes



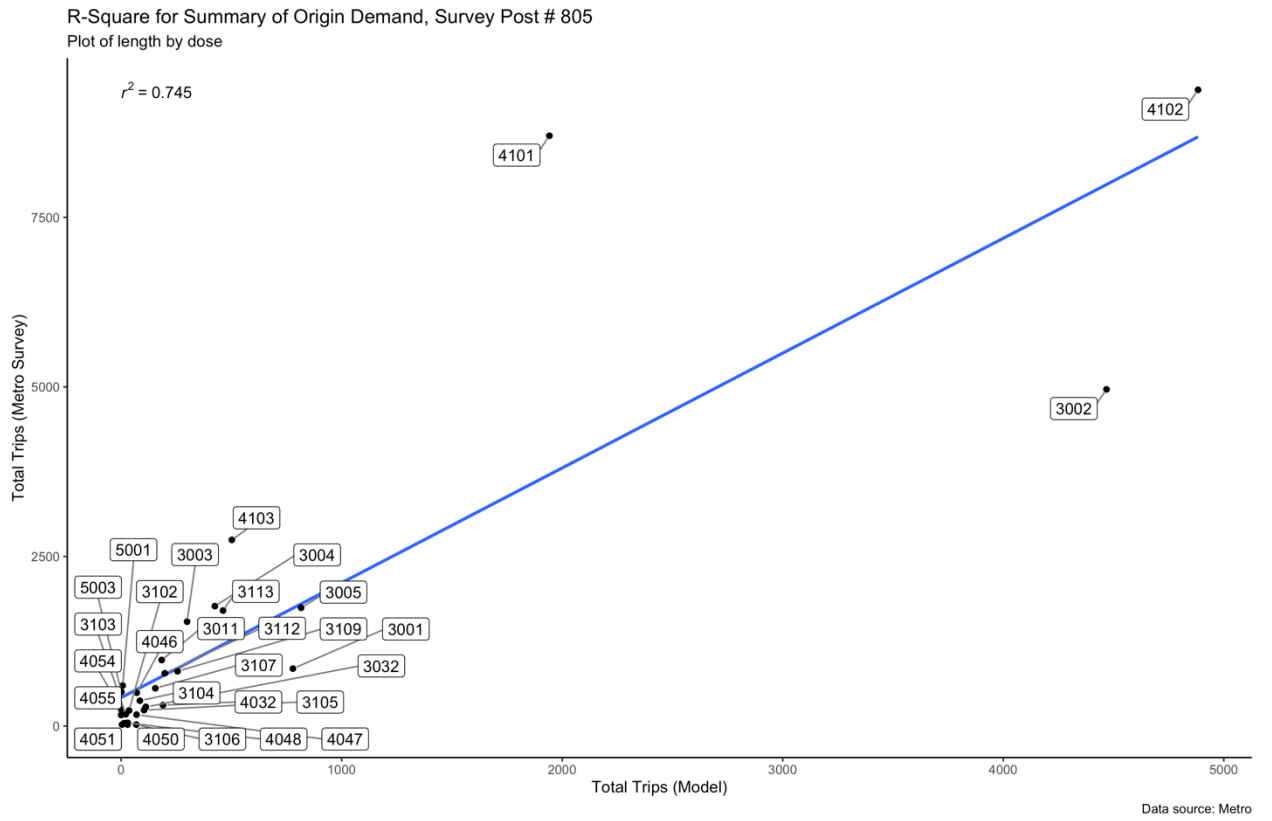
Source: Author

Figure 33: *Rodovia Anchieta*, trip distribution, zone extrapolation

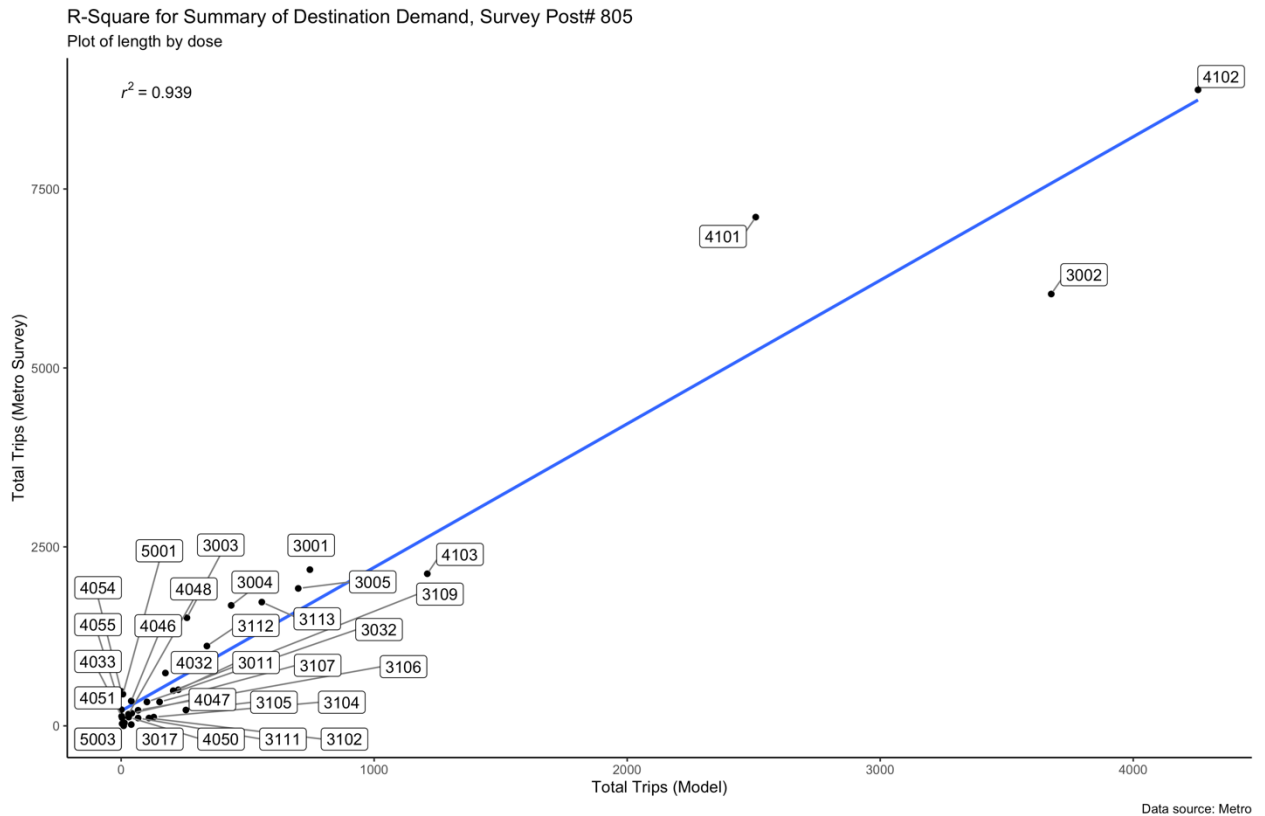
Source: Author

Afterward, we could plot the model outputs in a sequence of images from the summarized perspective or origin and another from destination demand. For example, Figure 34 and Figure 35 shows the R-squared analysis and result for OD survey outpost #805, *Rodovia dos Imigrantes*. R-squared achieved was 0.745 and 0.939, respectively.

Figure 34: R-squared analysis of origin demand in outpost 805 (*Imigrantes*)



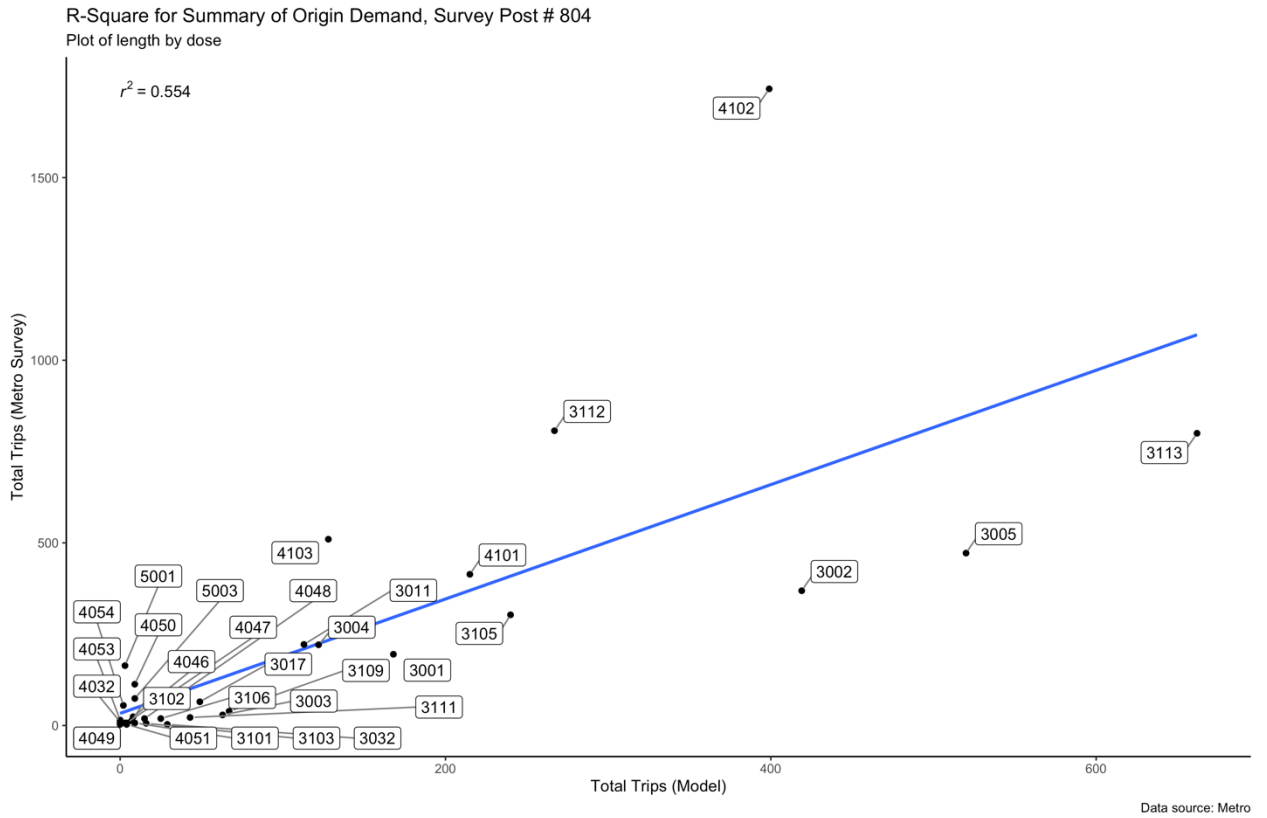
Source: Author

Figure 35: R-squared analysis of destination demand in outpost 805 (*Imigrantes*)

Source: Author

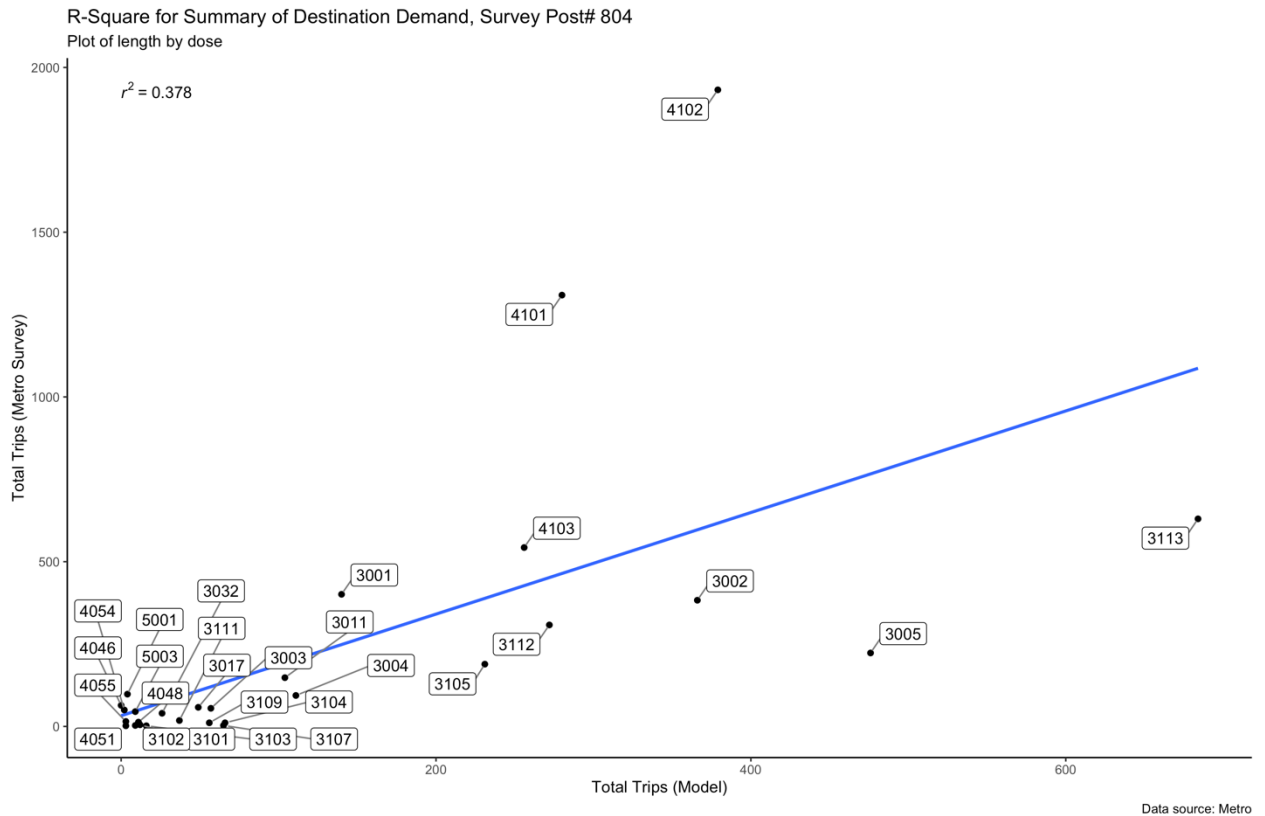
The results for *Rodovia dos Imigrantes* were within this dissertation's expectations and one that had prospects of being more accurate since Santos and São Vicente acts by far as the most attractive regions for SAI. Figure 36 and Figure 37 show the R-squared analysis and OD survey outpost #804, *Rodovia Anchieta*. R-squared achieved was 0.554 and 0.378, respectively.

Figure 36: R-squared analysis of origin demand in outpost 804 (Anchieta)



Source: Author

Figure 37: R-squared analysis of destination demand in outpost 804 (Anchieta)



Source: Author

The results for *Rodovia Anchieta* were under this dissertation's expectations and could not be explained by a lack of sufficient data. Although, this was well within expectation since the transportation profile of *Rodovia Anchieta* is one of commercial and heavily truck-based modes of transportation. R-squared plots show that the model supplied flows below the expected amount for especially zones 4102, Santos, a destination that has a much stronger influence in commercial trips than passenger demand, due to the presence of the Port of Santos. As explained in previous sections, commercial transportation was not contemplated in this dissertation.

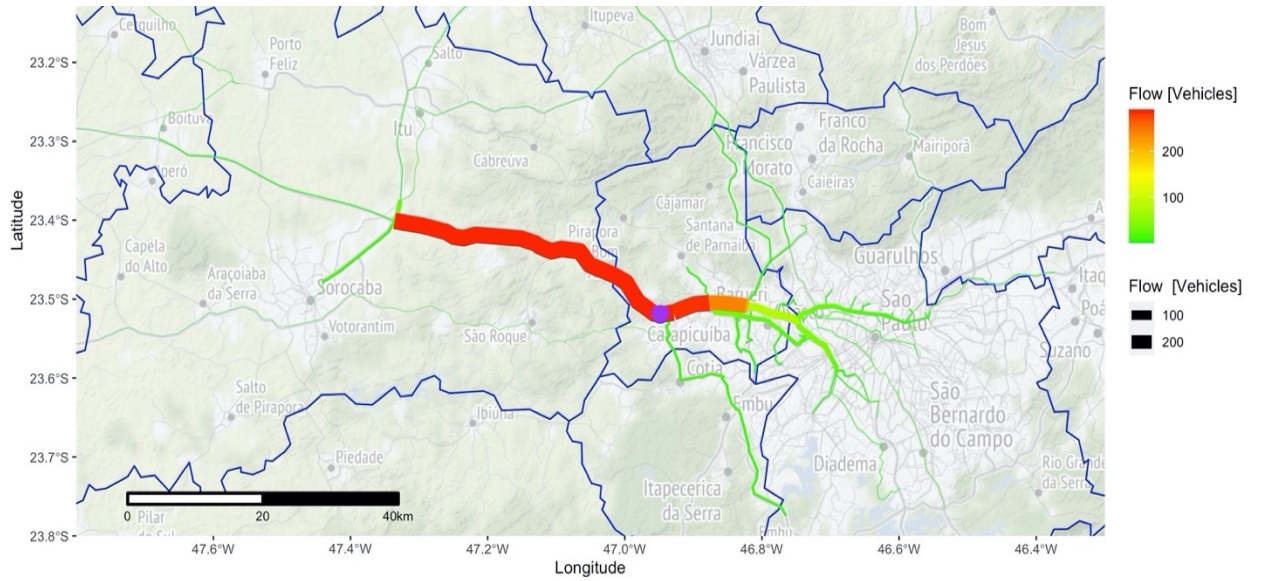
### ***Sistema Castelo-Raposo, São Paulo – Sorocaba***

The *Presidente Castelo Branco* highway, together with *Raposo Tavares*, also recognized as the *Sistema Castelo-Raposo*, is the main interconnection between the metropolitan region of São Paulo and the western-central region, as well as access to Argentina and Paraguay. In addition, the system is also known for heavy traffic between the municipalities of São Paulo and Osasco. Following Figure 38 and Figure



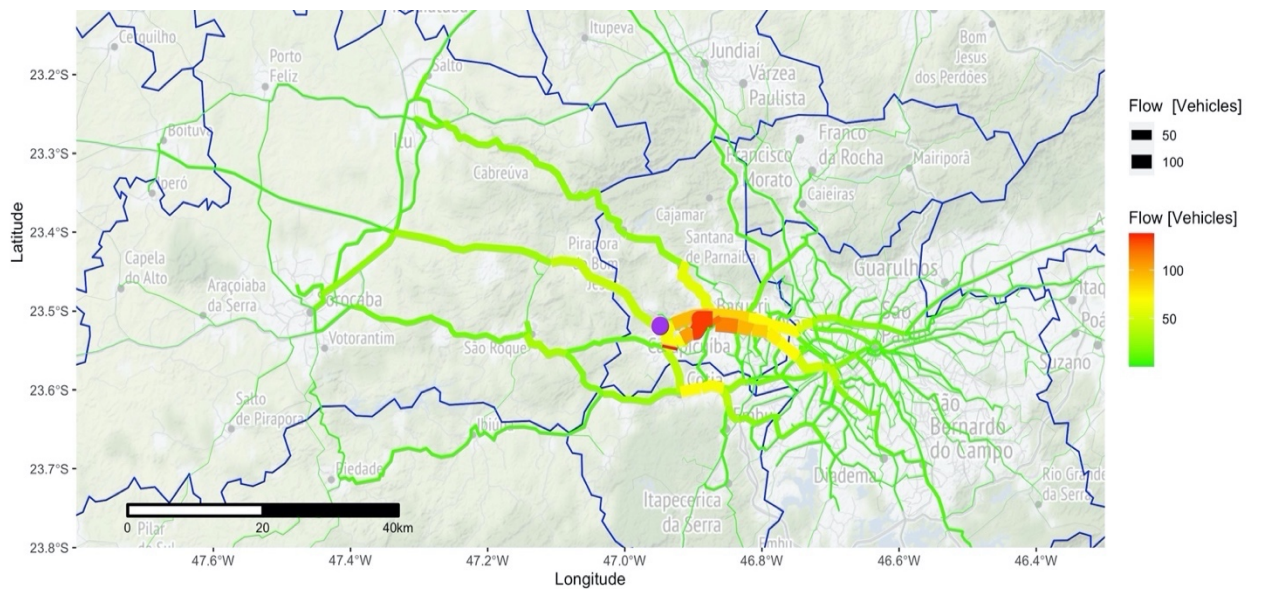
39 show the assignment result before and after the trip endpoint estimation algorithm for *Rodovia Presidente Castelo Branco*.

Figure 38: *Rodovia Presidente Castelo Branco*, trip distribution, partial routes



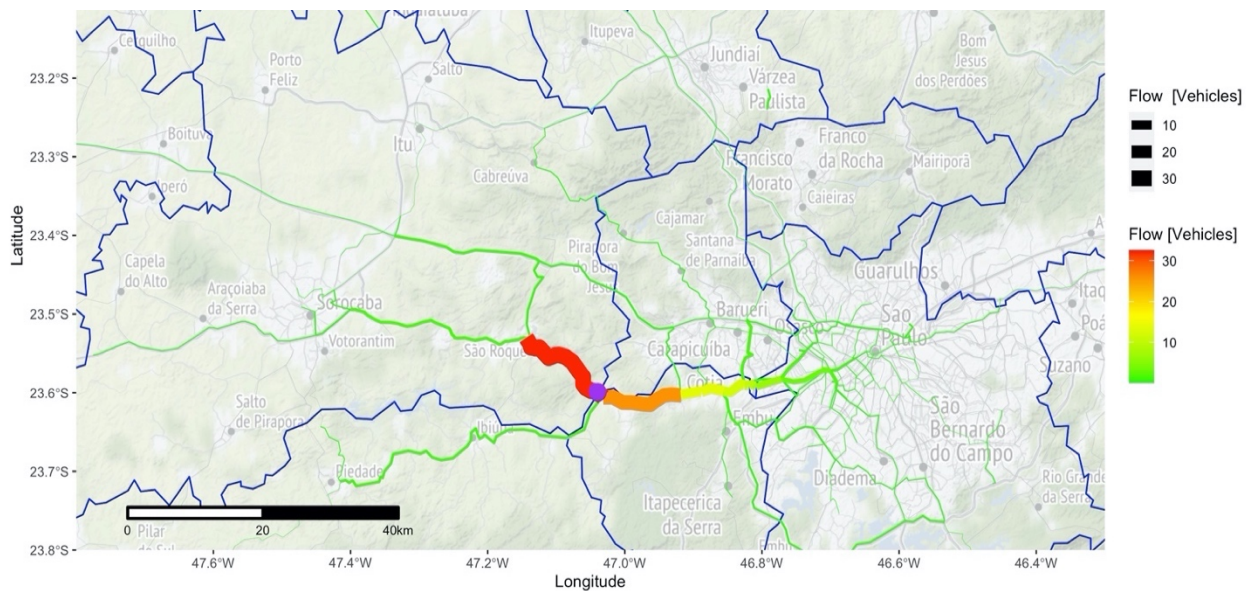
Source: Author

Figure 39: *Rodovia Presidente Castelo Branco*, trip distribution, zone extrapolation

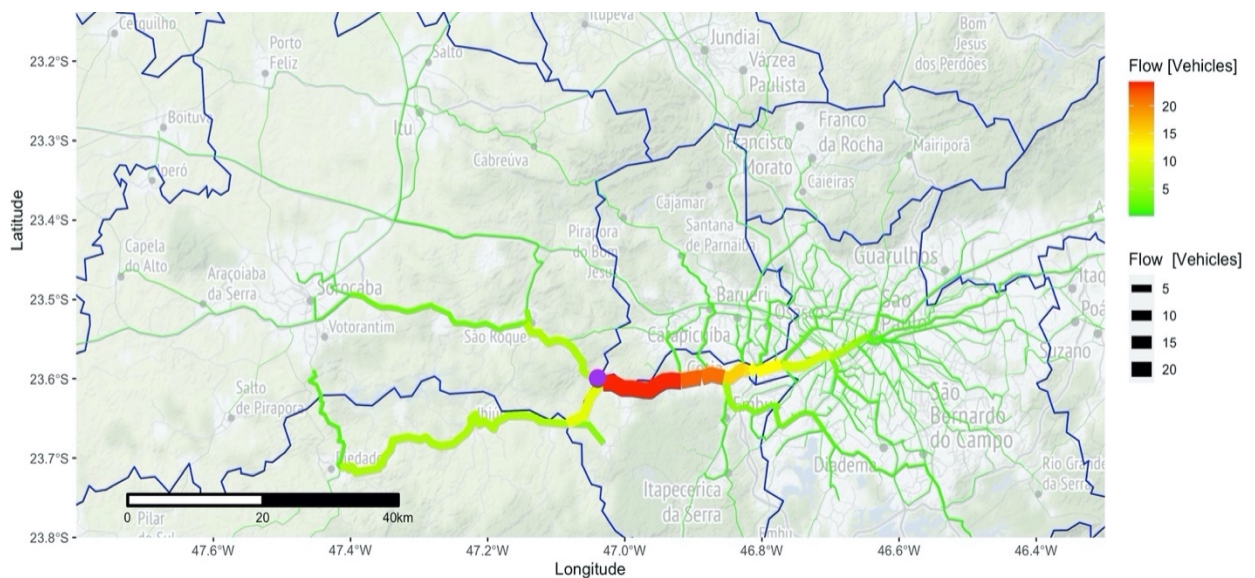


Source: Author

Figure 40 and Figure 41 show the result of the assignment prior to and after the trip endpoint estimation algorithm for *Rodovia Raposo Tavares*.

Figure 40: *Rodovia Raposto Tavares*, trip distribution, partial routes

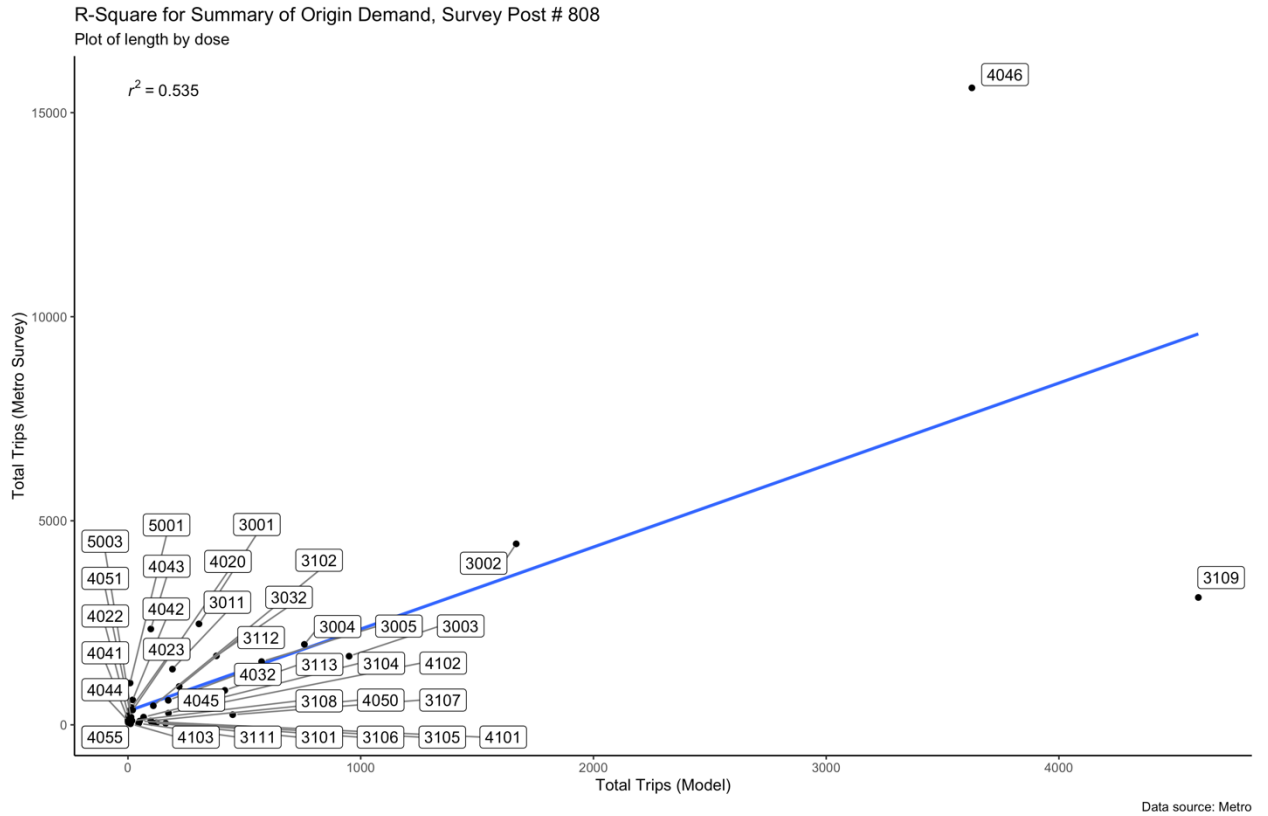
Source: Author

Figure 41: *Rodovia Raposto Tavares*, trip distribution, zone extrapolation

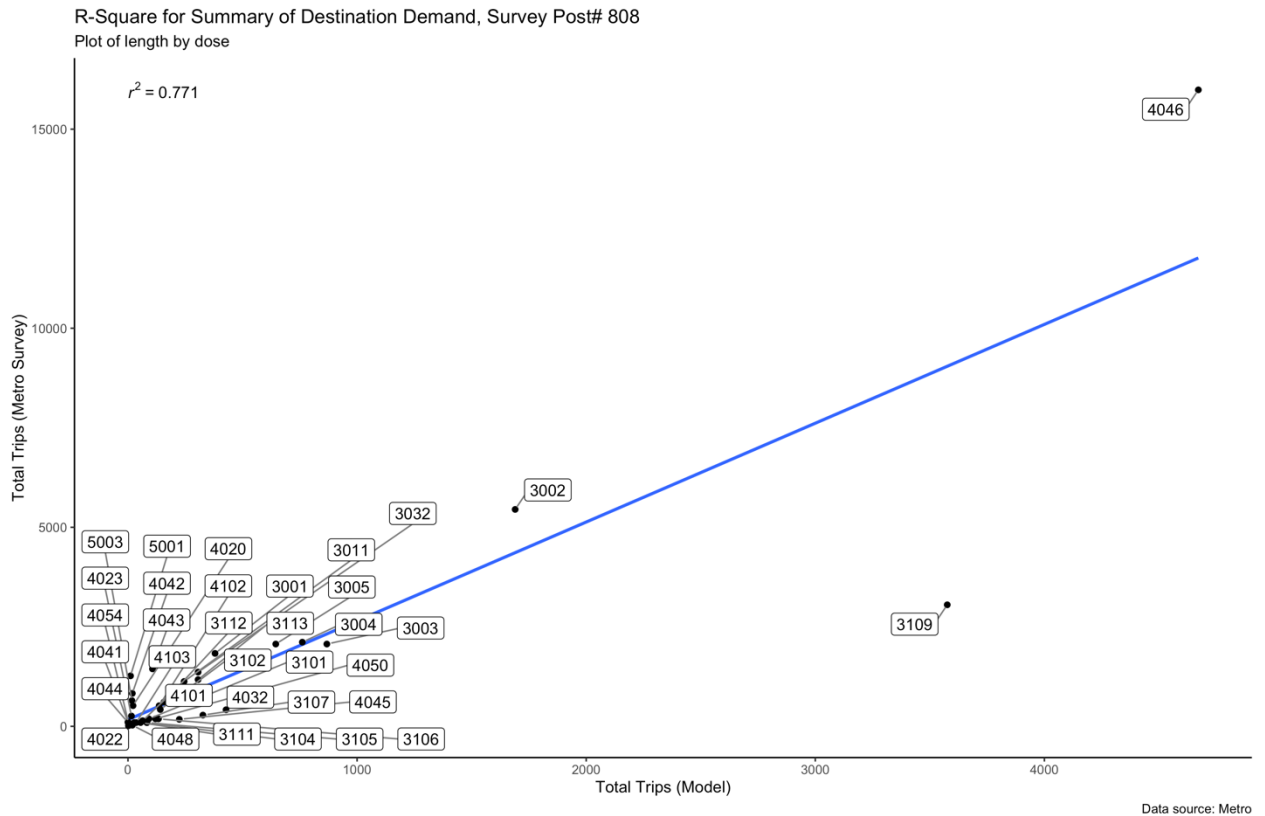
Source: Author

Afterward, we could plot the model outputs in a sequence of images from the summarized perspective or origin and another from destination demand. Figure 42 and Figure 43 shows the R-squared analysis and result for OD survey outpost #808, *Rodovia Presidente Castelo Branco*. R-squared achieved was 0.545 and 0.771, respectively.

Figure 42: R-squared analysis of origin demand in outpost 808 (*Presidente Castelo Branco*)



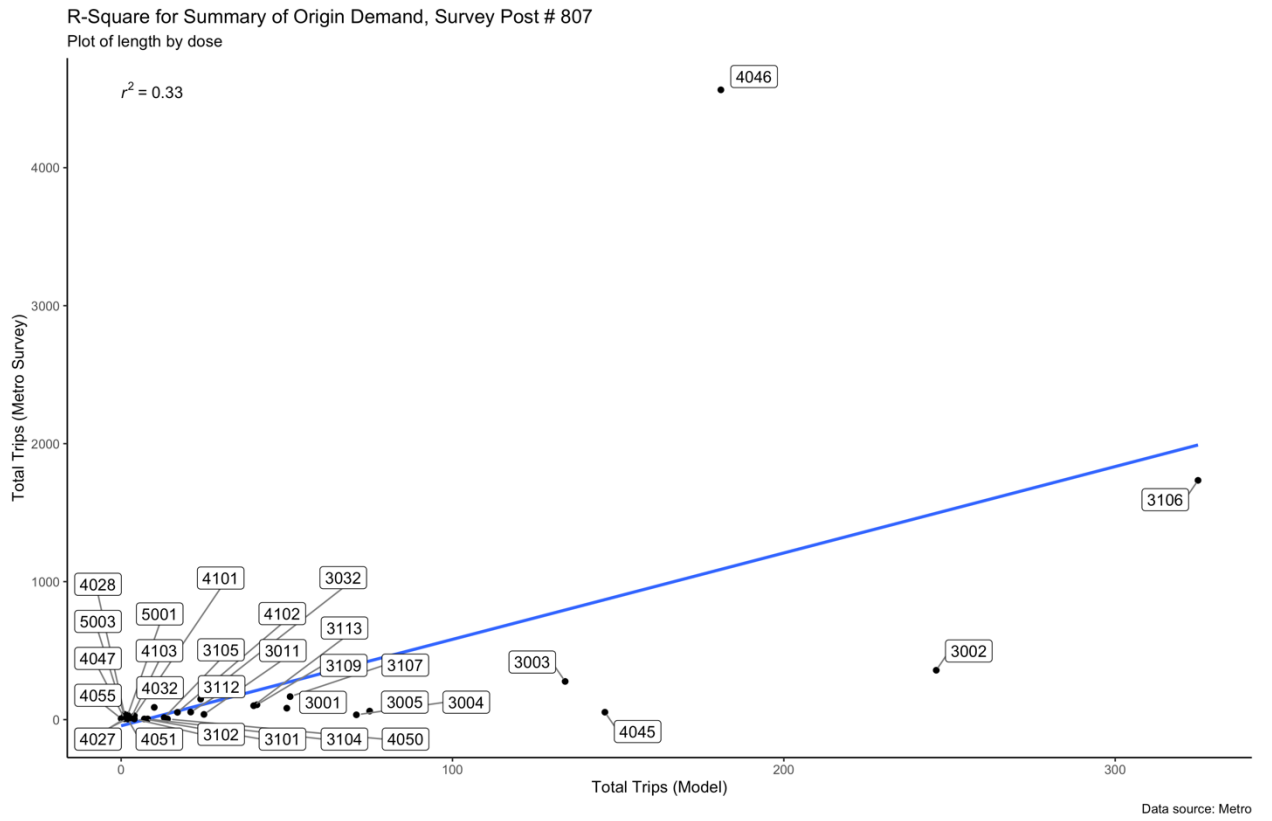
Source: Author

Figure 43: R-squared analysis of destination demand in outpost 808 (*Presidente Castelo Branco*)

Source: Author

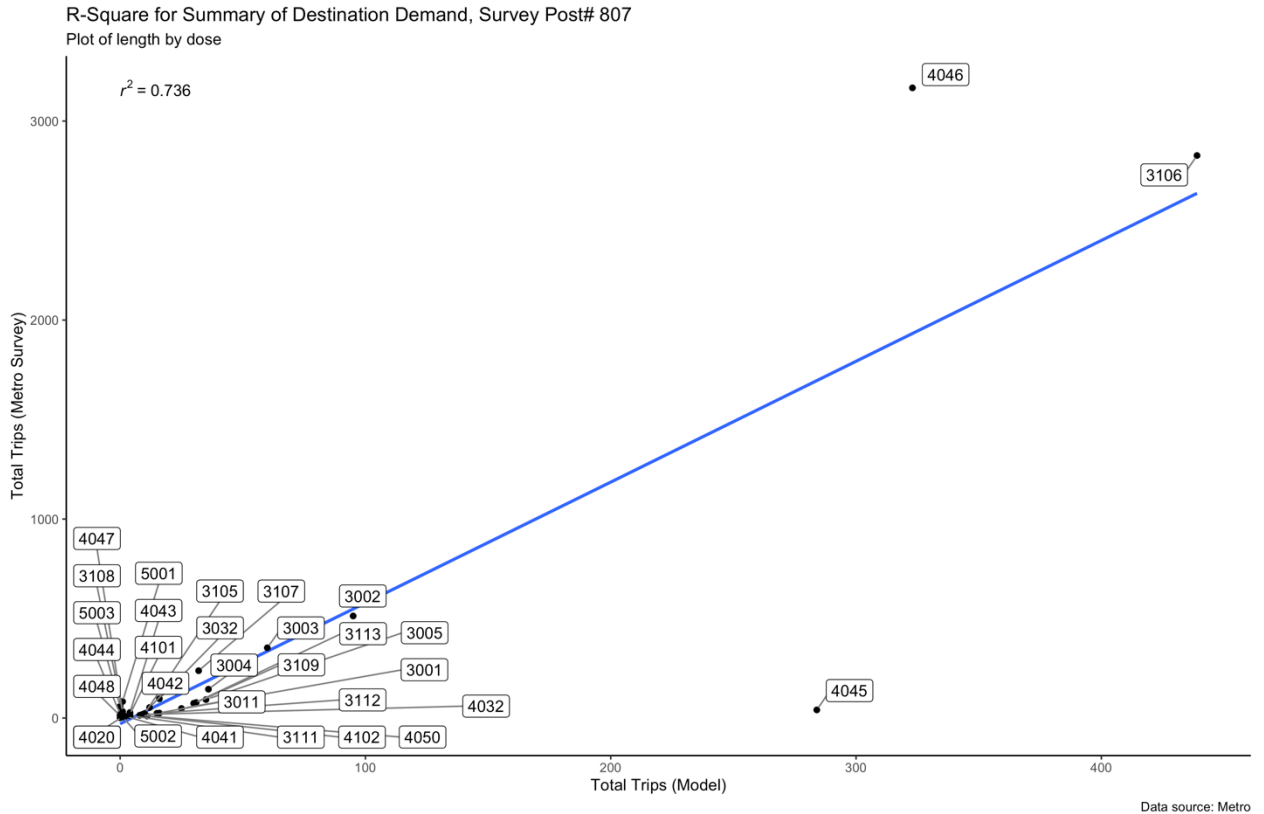
The results for *Rodovia dos Presidente Castelo Branco* were within this dissertation's expectations and could mostly be improved by further detailing the network and zones and a more robust approach to model attraction vectors. Figure 44 and Figure 45 show the R-squared analysis and result for OD survey outpost #807, *Rodovia Raposo Tavares*. R-squared achieved was 0.33 and 0.736, respectively.

Figure 44: R-squared analysis of origin demand in outpost 807 (Raposo Tavares)



Source: Author

Figure 45: R-squared analysis of destination demand in outpost 807 (Raposo Tavares)



Source: Author

The results for *Rodovia Raposo Tavares* were under this dissertation's expectations and could not be explained by a lack of sufficient data. We concluded that toll avoidance in SP270-079 is a possible culprit since this tool has many possible escape routes considering its close to the municipality of *Alumínio*.

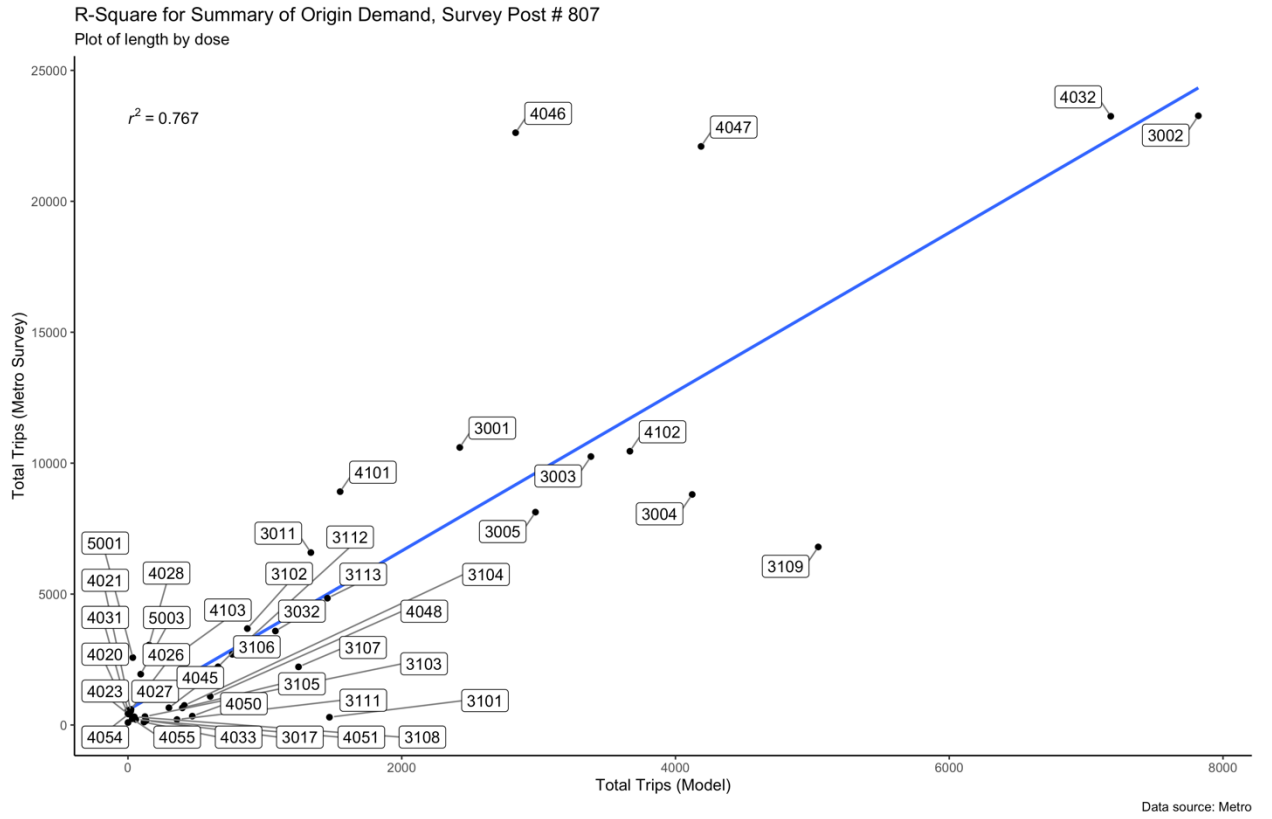
### ***Sistema Dutra-Ayrton Senna, São Paulo – São José dos Campos***

Unfortunately, due to the lack of access to data from the Dutra highway, this transportation system significantly reduced algorithm output quality, which is expected from such a significant deficiency of data. For these reasons, results were excluded, with R-squared results under 0.1.

### ***Aggregate Analysis***

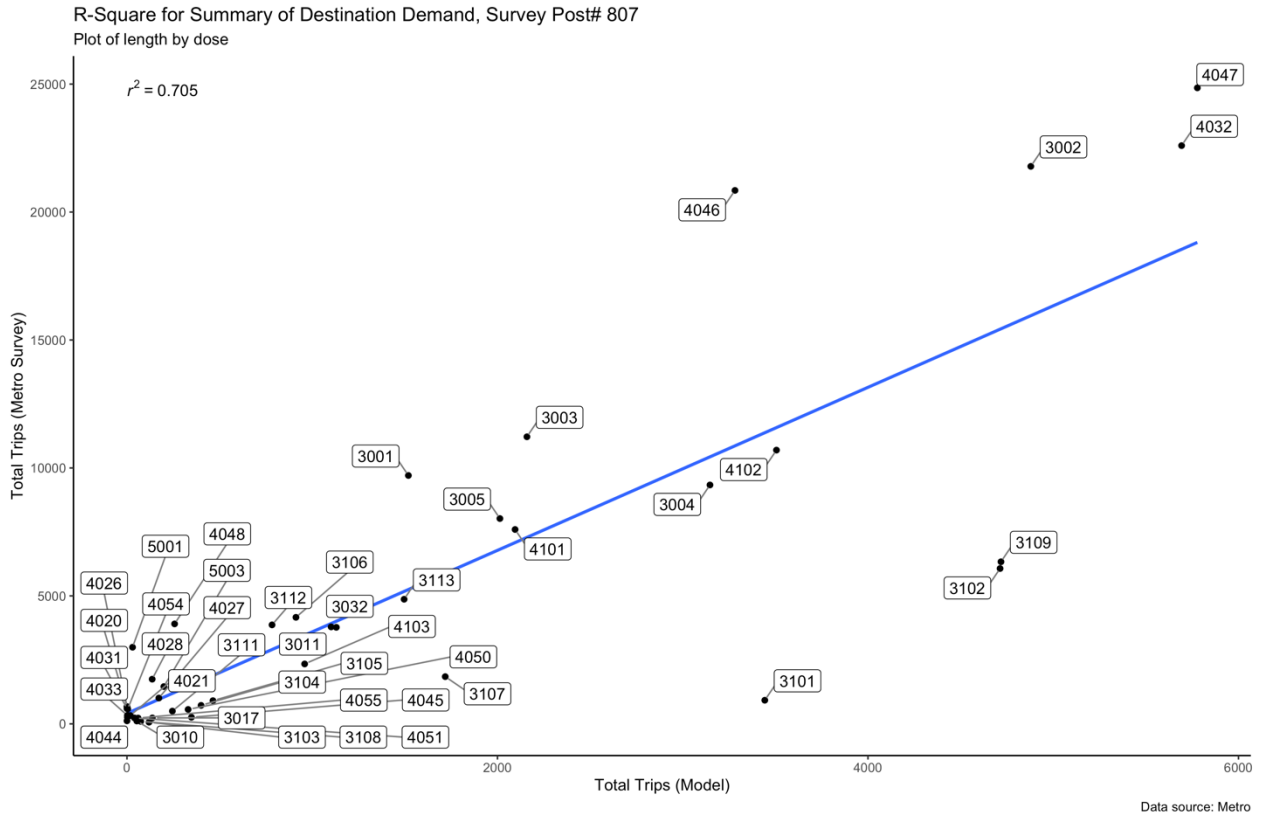
As a final analysis for R-squared between survey and model trip distribution data, an aggregate plot of all results is shown in Figure 46 and Figure 47. R-squared achieved was 0.767 and 0.705, respectively.

Figure 46: R-squared analysis of aggregate origin demand



Source: Author

Figure 47: R-squared analysis of aggregate destination demand



Source: Author

This section concluded the R-squared analysis and demonstrated a good approximation of the algorithm faced with real-world survey data, with a global R-squared of around 0.74.

### 5.3. Chapter final remarks

This chapter presents the proposed a method for validating the distribution of trips. The achieved R-squared metric demonstrated a good approximation of the algorithm faced with real world survey data, with a global R-squared of around 0.74 showing a suitable estimate of real-world profiles in intercity traffic. Table 5 displays the results from the individual analysis locations in the aggregate analysis. A more robust and well-grounded technical approach to estimating attraction vectors could yield great improvements in the OD distribution analysis, consisting a subject for further study.



Table 5: Individual location R-squared analysis

Location	Survey ID	Origin R-squared	Destination R-squared	Comments/ Issues
Anhanguera	809	0.463	0.477	This region requires additional network details and a more in-depth modeling of attraction vectors due to high amounts of urban traffic in this section of the highway.
Bandeirantes	810	0.889	0.591	
Imigrantes	805	0.745	0.939	
Anchieta	804	0.554	0.378	Issued due to mainly commercial vehicle usage attributed to the Port of Santos.
Castelo Branco	808	0.545	0.771	
Raposo Tavares	807	0.33	0.736	Issues due to high % of urban traffic among intermunicipal
Dutra-Ayrton	* Insufficient AVI data resulted in R-squared under 0.2			

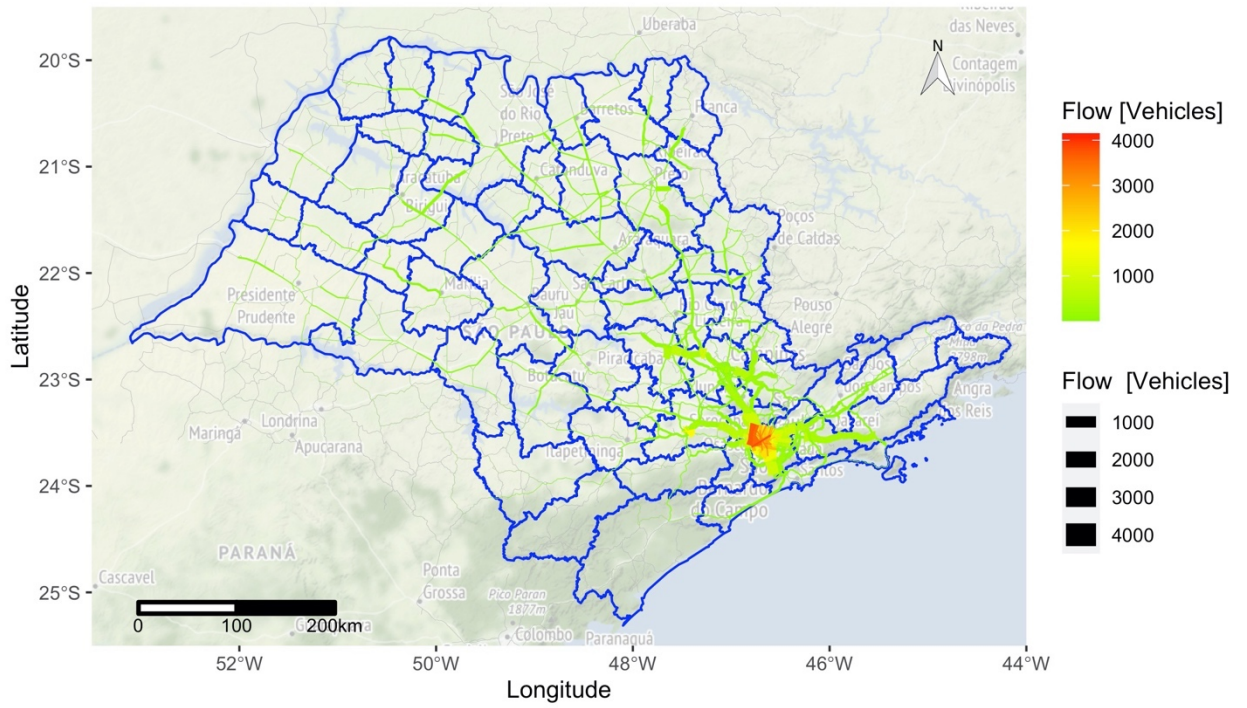
## **6. Results**

This chapter presents the estimated OD matrix from the combined activity groups. It examines the result with a selection of methods to give a verdict on the attempt of matrix estimation, its applicability, and suggestions in techniques to improve its results.

### **6.1. Partial Route reconstruction results (Step 1)**

For this section, to analyze the result of the partial route reconstruction, a matrix was assembled. By grouping each partial route for each equal origin and destination, all cells of the OD matrix are defined. Figure 48 shows a map representing the assignment of partial routes of 1,931,561 trips recognized for 22/03/2017, based on 1,230 AVI equipment, part of them on state highways (144 ETC equipment) 1,086 radars. These allocated trips exclude 2% expurgated trips that have been filtered out, applying the eligibility criteria. The identification of partial routes was performed using the procedure described in section 4.1. There is a precise concentration of trips around Sao Paulo state's most densely populated zones.

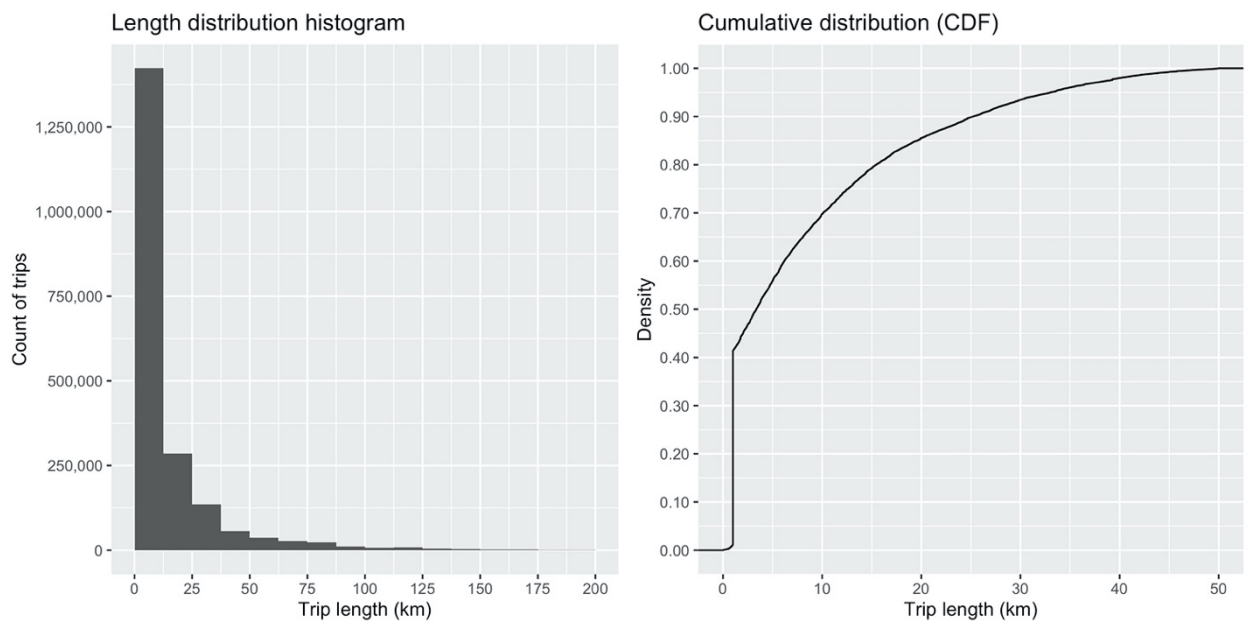
Figure 48: Partial route matrix assignment



Source: Author

Figure 49 displays the trip length distribution and the cumulative distribution for Step 1.

Figure 49: Analysis of partial route matrix (Step 1)

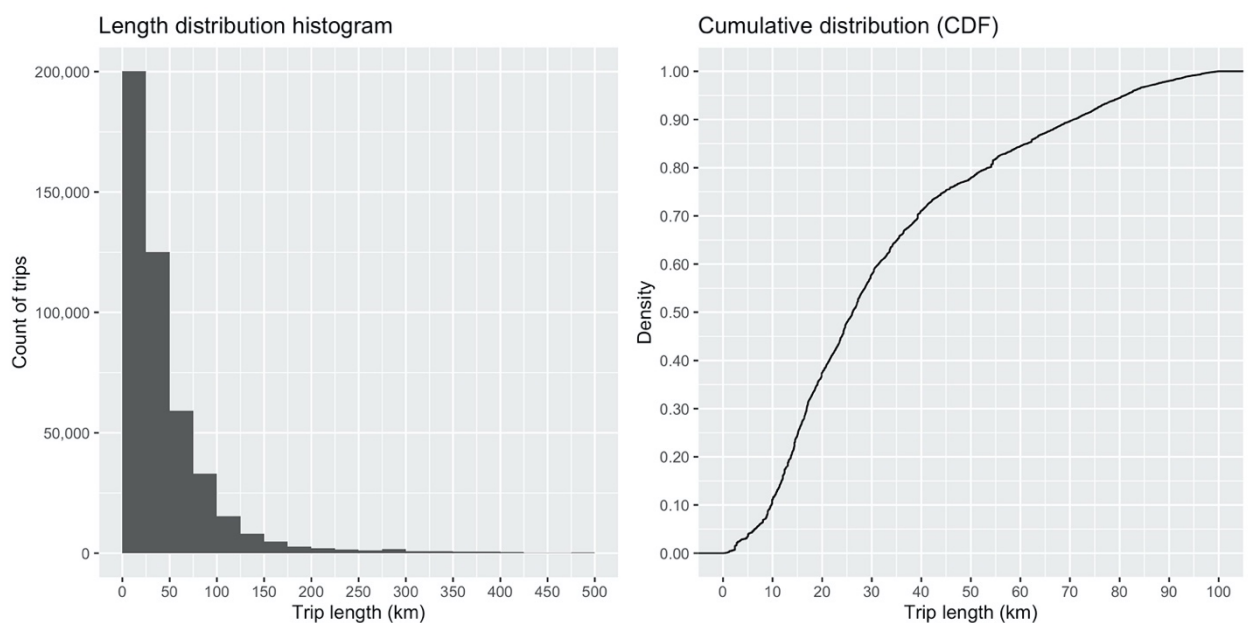


Source: Author

The average trip length for this dataset is 15km, intracity trip length at 10km, and intercity traffic at around 46km. For the sampled period, the most extended trip recorded was 772km.

Figure 50 shows the distribution of intercity trips. An essential aspect of this result is that single passages on ETC locations are recorded as 0km length trips (as seen on the vertical line in Figure 49), which would not contribute to intercity traffic. Without considering 0km length trips, intercity traffic is around 22%. Considering that trips that pass through ETCs and considering the distribution of equipment in the state of São Paulo are likely to be intercity trips when grouping these 0km length trips into intercity traffic, we reach a more realistic 48% for intercity trips. Both the interpretation and consideration of 0km length trips has been a topic of much discussion in this dissertation, while potentially bringing insights into a greater amount of passenger movement patterns, they can also cast doubt in its extrapolation of origin and destinations due to the lack of constraints in their determination. Further studies could bring insights in the value of usage or elimination of these trips, as well as propose modified trip origin and destination extrapolation.

Figure 50: Analysis of partial route matrix (Step 1) for intercity traffic

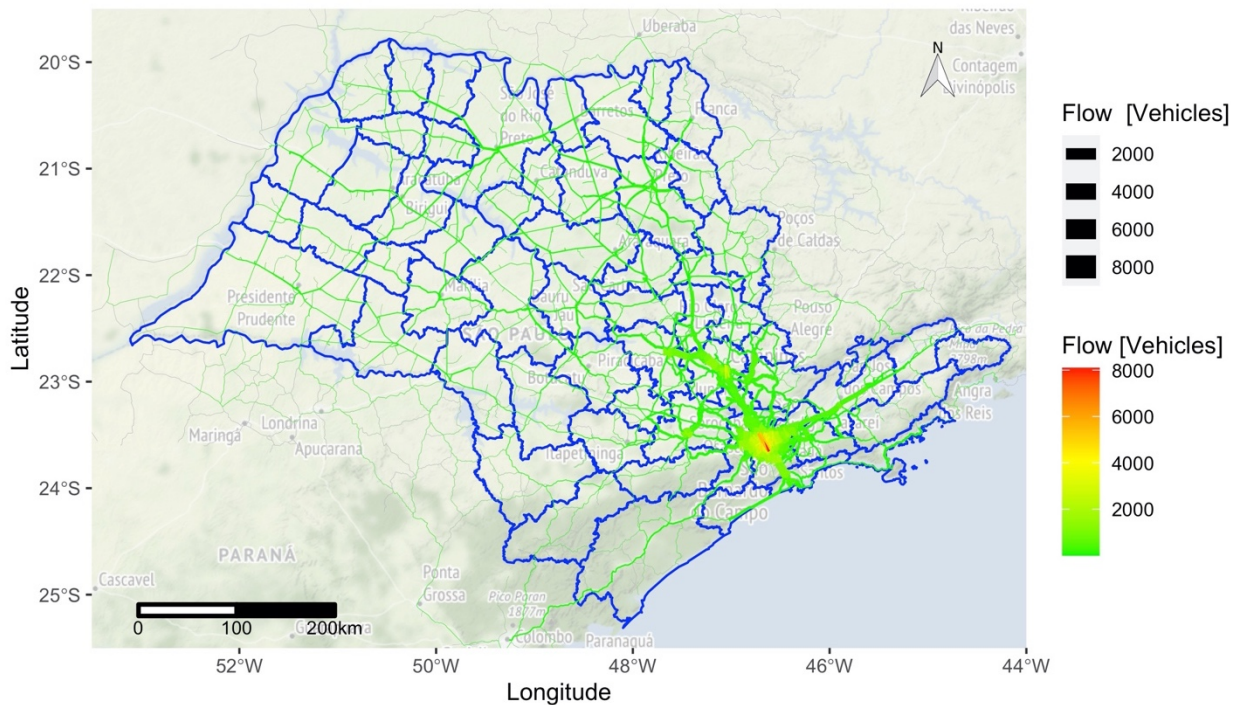


Source: Author

## 6.2. O/D Matrix construction results (Step 2)

The resulting extrapolated routes matrix, generated with the application of the algorithm to the partial route matrix, is shown in Figure 51, with a total of 1,931,561 unique trips for one day.

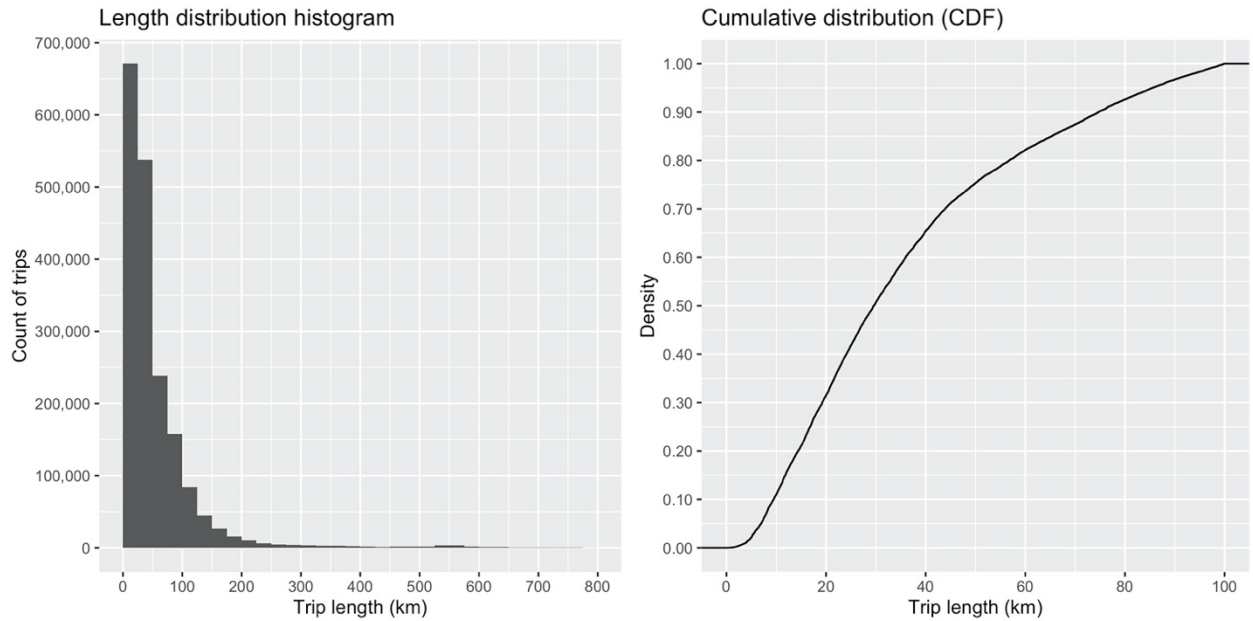
Figure 51: O/D Matrix construction (Step 2) assignment



Source: Author

Matrix assignment of this initial step shows a promising first glance at the profile of intercity traffic in the State of São Paulo, and even exhibiting a suitable approximation from intracity traffic in the city of São Paulo, with its several OCR equipment placements. Figure 52 displays the trip length distribution and the cumulative distribution for Step 2.

Figure 52: Analysis of O/D Matrix construction (Step 2)

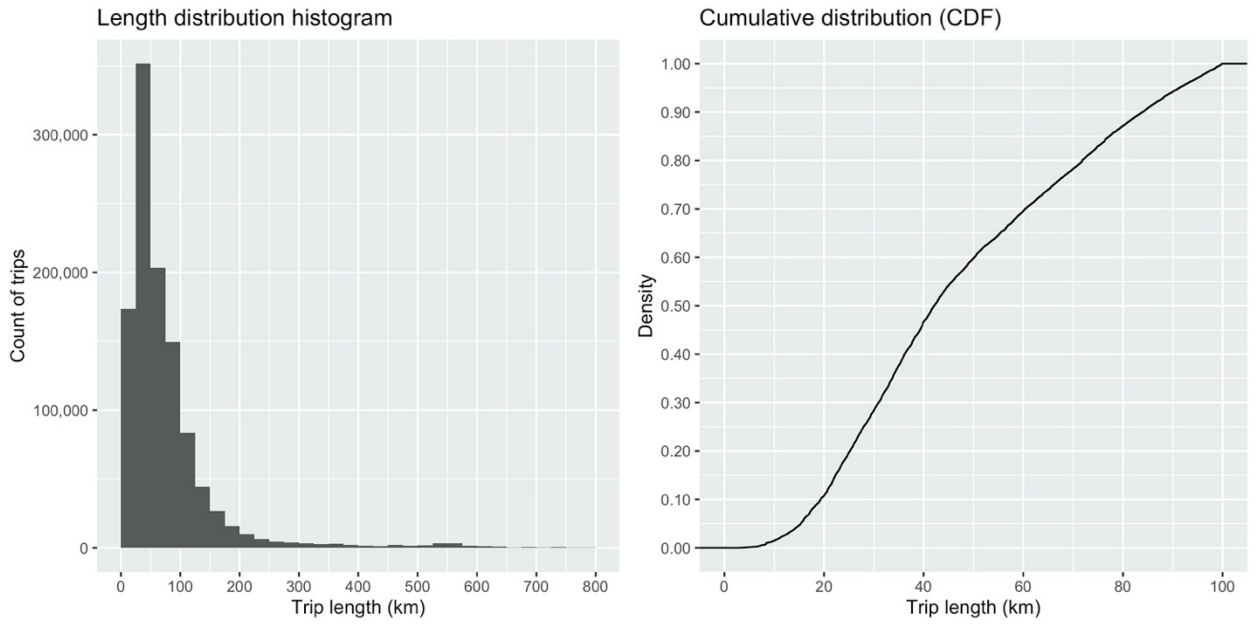


Source: Author

The analysis of the resulting extrapolated matrix demonstrates a significant increase in trip length. After extrapolation of origin and destinations for the previously described 0km trips, resulting trip lengths contributed to this increase in overall length. Extrapolating 0km trips is a contested subject and while presented in this dissertation, would benefit from further analysis.

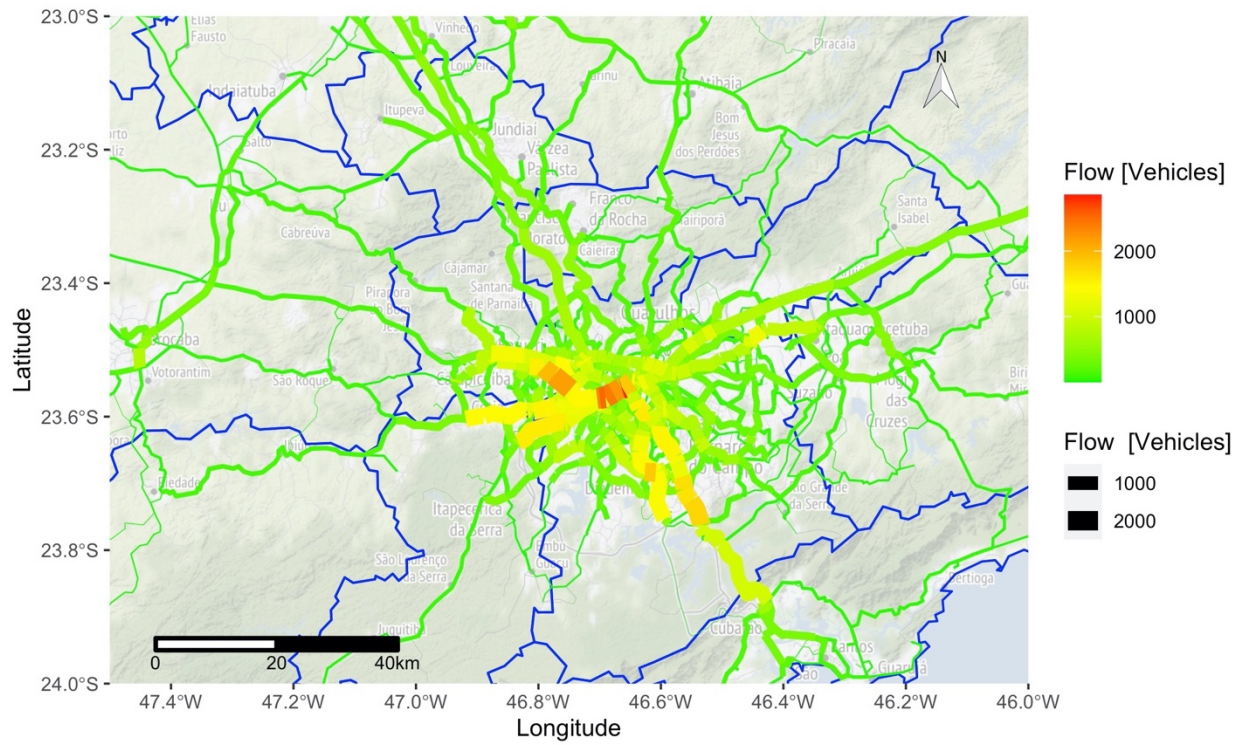
The average trip length for this dataset is 55km, intracity trip length at 21km, and intercity traffic at around 78km. The most extended trip generated was 1,200km and intercity traffic at around 60% for the sampled period, a 12-percentage point increase from the partial route analysis. Figure 53 displays the trip length and cumulative distribution for the intercity traffic for this step. The assignment of the intracity matrix is presented in Figure 54, showing a profile of traffic without noticeable flaws.

Figure 53: Analysis of O/D Matrix construction (Step 2) for intercity traffic



Source: Author

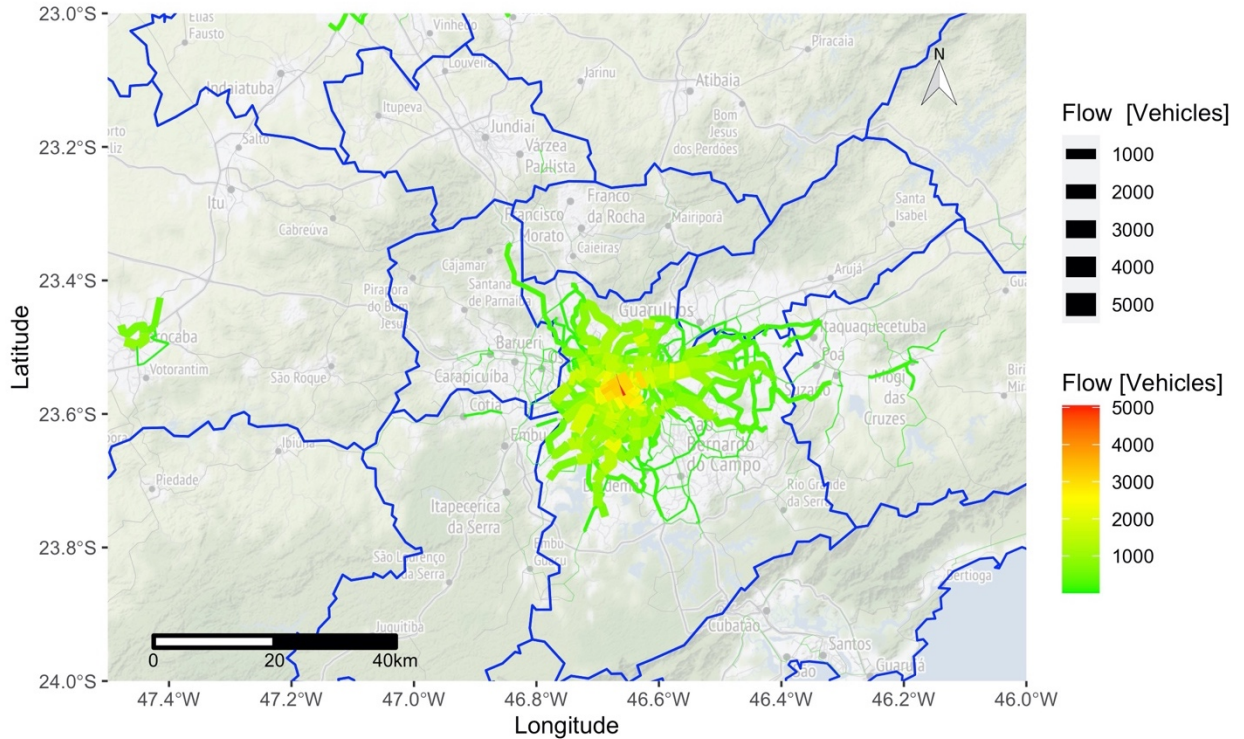
Figure 54: O/D Matrix construction (Step 2) without noticeable flaws



Source: Author

Figure 55 presents the assignment of intracity traffic around the metropolitan region of São Paulo (MRSP). It shows the trips within city bounds and respects intracity traffic characteristics.

Figure 55: Intracity O/D Matrix construction (Step 2) assignment for the MRSP



Source: Author

### 6.3. Matrix calibration results (Step 3)

Once the matrix estimation algorithm was applied, a seed matrix (Step 2) could be used to input the transportation planning software PTV Visum. Setting up a network, connectors, capacities, costs, and so on is an exhaustive process and not the focus of this dissertation, so that no further detail will be discussed.

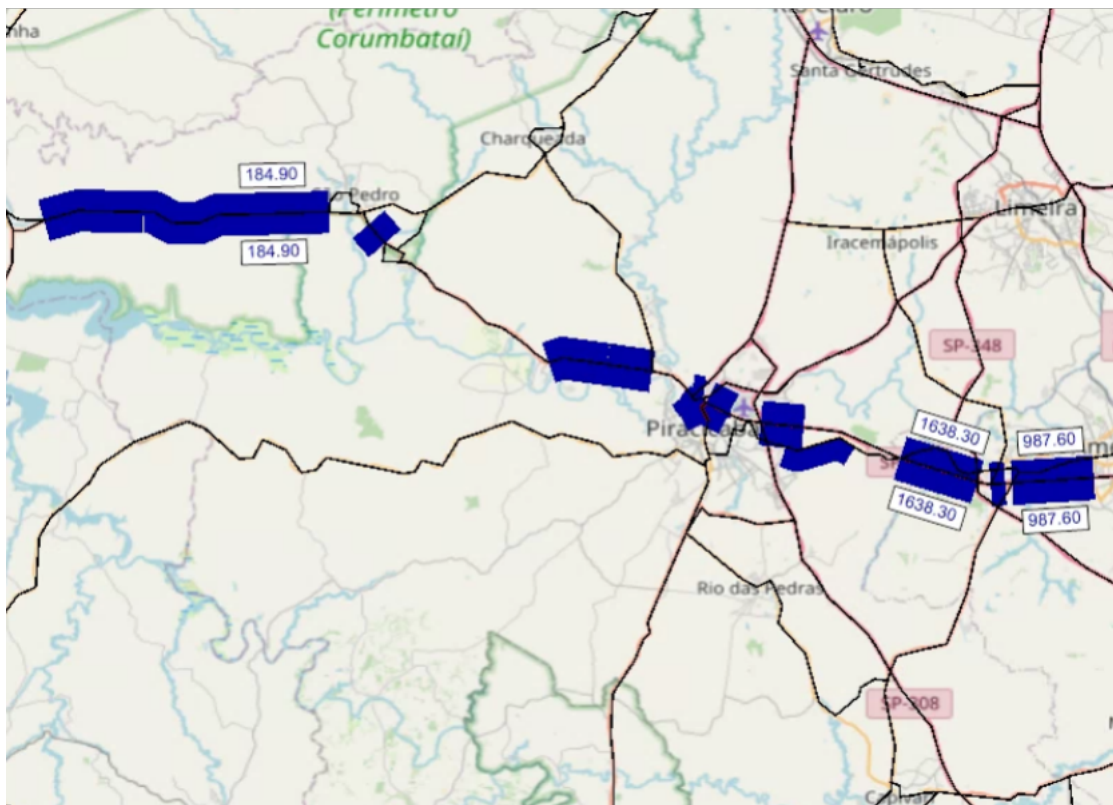
With the appropriate setup and data entered on PTV's Visum, we could create a selection of the count data that was used in the calibration algorithm T-Flow Fuzzy.



### 6.3.1. Selection of count data

The total number of count locations available was 1,192, which is considerably large compared to the number of zones. For this reason, it was necessary to select a more specific set of data points carefully. The first exclusion was locations with a high percentage of urban traffic, the reason being that this dissertation focuses on intercity traffic, resulting in 744 count locations. Then we removed count locations where the assigned amount was zero, carefully considering not removing count data relevant to this dissertation, such as dual carriageway high traffic roads. Lastly, we singled out points in more remote locations, not within proximity of high traffic roads, and those with multiple close concurrent locations, such as those in Figure 56. These multiple count locations within proximity can disturb the calibration algorithm, having different targets for the same road segment when there is not any zone that could attribute the variation in traffic volume.

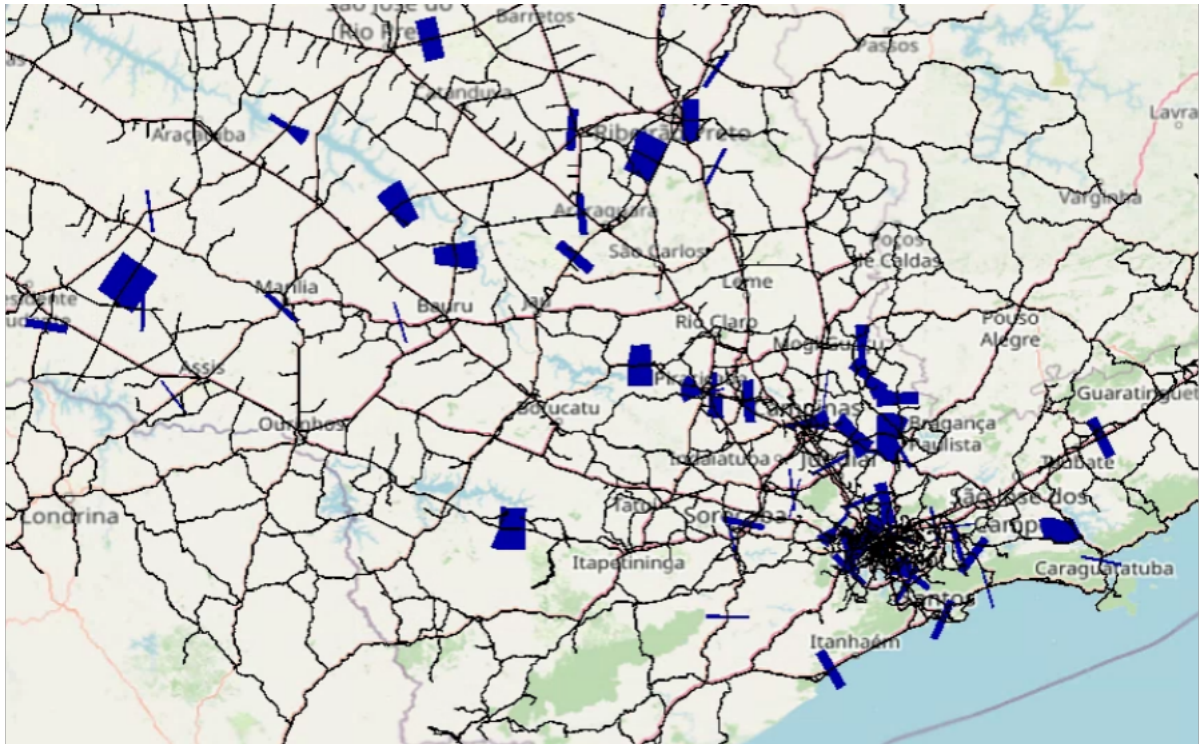
Figure 56: Count data points with close proximity



Source: (Costa, Breno, 2021)

The final selected set of count locations is presented in Figure 57.

Figure 57: Final count data location selected



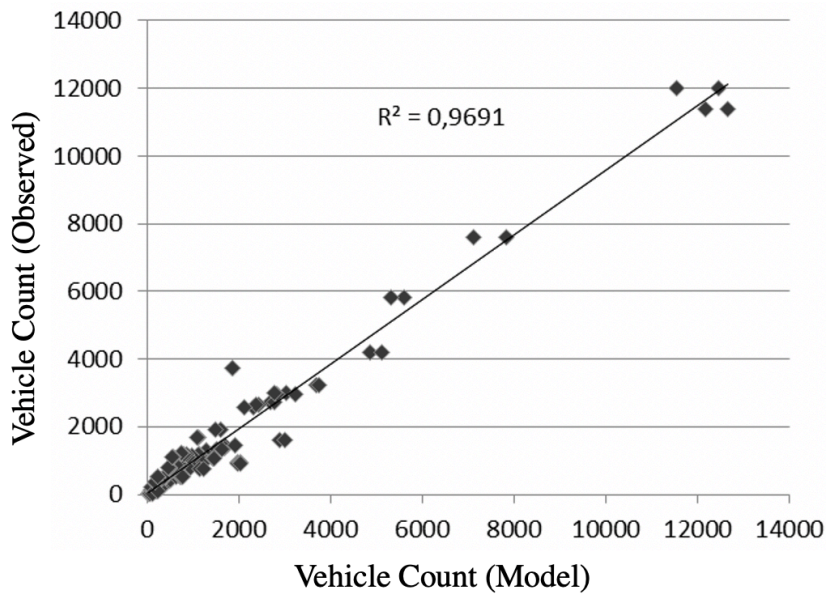
Source: (Costa, Breno, 2021)

### 6.3.2. T-Flow Fuzzy algorithm results

Once completed the selection of points, we set the necessary configurations within the VISUM and proceeded to run with the set amount of count data as inputs in the calibration algorithm. Following the algorithm's output, with results tabulated, we plotted an R-squared of observed and model traffic shown in Figure 58.

The algorithm proposed supplied the calibration algorithm with a quality seed matrix for count volume correction, achieving an  $R^2$  of over 0.96. The coefficient  $R^2$  indicates how well the regression predictions approximate the real points. The value found (0.96) indicates that the estimated data provides very accurate results compared to the real flow of vehicles measured in traffic counters.

Figure 58: R-squared of Model x Observed traffic



Source: Author

Being of standard practice in transportation engineering to analyze the GEH statistic (Equation 6), an established measure to evaluate how well the model represents real-world traffic. M represents the model traffic, and C, the observed traffic.

$$GEH = \sqrt{\frac{2(M - C)^2}{M + C}} \quad (6)$$

In more complex network models, a standard recommendation is to aim for 85% of calibration points under a GEH of 5. However, a value of 10 is also suitable (Friedrich et al., 2019). Friedrich et al. (2019) recommends the categories shown in Table 6.

Table 6: Friedrich et al. GEH category recommendations

SQV statistic	GEH	Evaluation
0.90	3.4 to 3.6	Excellent match
0.85	5.4 to 5.8	Good match
0.80	7.5 to 8.5	Acceptable match

	(Since the GEH statistic is not symmetrical, the same absolute deviation of a measured value upwards and downwards are evaluated differently)	
--	---	--

Our results show that 65% of the counting points provided GEH below 5, while 88% of them were below 10 (Table 7).

Table 7: T-Flow Fuzzy calibration GEH results

Average GEH	5.0
GEH < 5	65%
GEH < 10	88%

#### 6.4. Chapter final remarks

This chapter presented the results and analysis of the algorithm proposed in this dissertation. The partial route trip algorithm shows its capability in generating partial route reconstruction. An analysis of assignment and trip lengths showed a capability to extend trip lengths beyond first and last sensor detection. Most importantly the algorithm proposed supplied the calibration algorithm with a quality seed matrix for count volume correction, achieving an R-squared of over 0.9 and suitable GEH metrics with 88% of count sections showing a GEH under the value of 10.

The algorithm has trouble dealing with missing data from high traffic demand roadways, but besides data issues availability, the region in the study showed sufficient coverage to create traffic profiles for intercity traffic. Throughout the development of this research, new and improved methods advanced in popularity, with machine learning techniques becoming more mainstream, as such that the method proposed in this paper might have improved alternatives.

Some key aspects remain as for what is required to make use of this methodology: First an adequate network, zoning, and link cost determination. Second is AVI data sources, either from ETC, OCR, or as the case in this paper, both. The third is a large

collection of count data, for the calibration algorithm present in PTV Visum. Lastly, a way to determine weights for the distribution of flows, either by calibrating attraction vectors or making use of approximations, such as the one in this paper. As for validating the results, sampling of OD patterns in key locations should be implemented, either by surveys or other location tagging information, such as mobile location data, GPS sources, or other vehicle tracking technologies.

The state of São Paulo has the highest coverage of AVI technology out of the entirety of the country of Brazil, as such for the application on other regions, as well as regions in other countries with reduced AVI coverage should be proceeded with caution. For the scenarios in which researchers aim to apply the methodology or a variation of it, for regions that lack ETC information (information key to filtering candidate zones for extrapolation), a modification of the method can be implemented. A recommendation is to apply a layered approach to vehicles crossing OCR equipment, for example, considering the accuracy of detection in OCR equipment being around 80%, there is a 20% chance of it crossing the road section and not being recorded. For a second OCR equipment a vehicle that escaped detection, the probability of not being detected again decreases to 4%, as for the third failure of detection has a negligible probability of not being detected at under 1%. Each further OCR equipment barrier defines a layer of the network, with probabilities of paths ending on each layer according to the combined detection probability.

## 7. Conclusion

This dissertation proposes a novel approach to estimating OD matrices based on data from AVI and traffic count devices in large-scale road networks. The proposed method: a) builds vehicle trajectories; b) extrapolates the OD zones from partial routes, and c) distributes flows through the gravity flow distribution model.

We applied the method in the road network of the State of São Paulo, Brazil. This method was validated by contrasting the results with the data from the demand survey in the RMSP. As a result, the estimated OD matrix can be considered suitable for practical application, as the calibration reaches an accuracy ( $R^2$ ) above 0.96 and GEH below 10 in 88% of the calibration points.

The method provided suitable accuracy in the estimations when comparing the flows over a few network links through a regression modeling measure of performance and the GEH index. However, missing data is essential to the algorithm's success, though the studied region showed sufficient coverage to create traffic profiles for intercity traffic. The importance of modeling attraction for each zone with more detail and constructing a network cost calculation that accurately reflects the real world should be noted.

The OD matrix would benefit other studies to analyze and compare OD demand matrices from other resources and apply both into public transportation initiatives, such as intercity bus lines and passenger rail systems, and study road concessions, such as expected revenue capacity expansions, among other predictive estimations. Additionally, further studies can improve on this method by including capacity-restricted assignment in the trip-distribution phase, include trip distribution for commercial modes of transportation, by means of appropriate modelling of commercial attraction variables, as well as simplify commercial impact by including a capacity restriction based on a preload of commercial vehicle volume on network links. Capacity restriction is a method of adding fixed link loads (preload) appropriate in representing common commercial vehicle network usage.

## References

- ORTÚZAR S.; J. D. D., & WILLUMSEN, L. G. **Modelling transport**. Chichester, West Sussex, England, Wiley, 1990
- WILSON AG. **Forecasting Planning**. Urban Studies, 1969; 6(3):347-367.
- J Z Nanne; B G Heydecker **Transportation Networks: Recent Methodological Advances**, 1998, Volume 319
- SHEFFI Y. **Urban transportation networks**, 1985. Englewood (NJ): Prentice-Hall;
- WARDROP, J.G. **Some theoretical aspects of road traffic research**, **Proceedings of the Institution of Civil Engineers**, Part II, volume 1, number 2, pages 325-378, 1952
- FRIEDRICH. M; PESTEL E.; SCHILLER C.; SIMON R. **Scalable GEH: A Quality Measure for Comparing Observed and Modeled Single Values in a Travel Demand Model Validation** Transportation Research Record: Journal of the Transportation Research Board. Issue 2673, No 4, April 2019, ISSN 0361-1981, pages 722–732
- BERNARDI, Eli **Automatic vehicle identification systems, especially the license plate recognition**, 2017.
- YOUSEFIKIA M; MAMDOOHI A. R.; MORIDPOUR S.; NORUZOLIAEE M. H. and MAHPOUR A. **A study on the generalized TFlowFuzzy O-D estimation**, 2013.
- NIGRO M.; ABDEL FATAH A.; CIPRIANI E.; COLOMBARONI C.; FUSCO G.; GEMMA A. Dynamic O-D demand estimation: application of SPSA AD-PI method in conjunction with different assignment strategies. **Journal of Advanced Transportation**, vol. 2018, Article ID 2085625, 18 pages, 2018.
- WILLIS A. E. and MAY A. D. Deriving origin-destination information from routinely collected traffic counts. **Research Report, Institute of Transportation Studies, University of California**, 1981.
- NIHAN N. L. Procedure for estimating freeway trip tables. **Journal of the Transportation Research Board**, no. 895, Paper No HS-035 368, 1982.
- VAN ZUYLEN H.; WILLUMSEN L. G. The most likely trip matrix estimated from traffic counts. **Transportation Research Part B**, 1980; 14(3):281–293
- MICHAEL B. The estimation of origin-destination matrices by constrained generalized least squares. **Transportation Research Part B**, 1991; 25(1):13–22.

- YANG H.; LIDA Y.; SASAKI T. The equilibrium-based origin-destination matrix estimation problem. **Transportation Research Part B**, 1994; 28(2):23–33
- CREMER M.; KELLER H. A new class of dynamic methods for the identification of origin-destination flows. **Transportation Research Part B**, 1987; 21:117–32.
- LIN P. W.; CHANGE G. L. A generalized model and solution algorithm for estimation of the dynamic freeway origin destination matrix. **Transportation Research Part B: Methodological**, vol. 41, no. 5, pp. 554–572, 2007.
- SHERALI H. D. and PARK T. Estimation of dynamic origin-destination trip tables for a general network. **Transportation Research Part B: Methodological**, vol. 35, no. 3, pp. 217–235, 2001.
- XIE C.; KOCKELMAN K. M.; WALLER S. T. A maximum entropy-least squares estimator for elastic origin-destination trip matrix estimation. **Procedia-Social and Behavioral Sciences**, vol. 17, pp. 189–212, 2011.
- CASTILLO E.; MENÉNDEZ J. M.; SÁNCHEZ-CAMBRONERO S.; CALVIÑO A.; and SARABIA J. M A hierarchical optimization problem: estimating traffic flow using gamma random variables in a bayesian context. **Computers & Operations Research**, vol. 41, no. 1, pp. 240–251, 2014.
- TOBIAS P. and BERNHARD F. A combined method to forecast and estimate traffic demand in urban networks. **Transportation Research Part C**, 2013; 31:131–144.
- JIANG D. D.; XU Z. Z.; XU H. W.; HAN Y.; CHEN Z. H.; YUAN Z. An approximation method of origin-destination flow traffic from link load counts. **Computer and Electrical Engineering**, 2011; 37:1106–1121.
- MUSSONE L.; GRANT M. S.; CHEN H. B. A neural network approach to motorway OD matrix estimation from loop counts. **Journal of Transportation Systems Engineering and Information Technology**, 2010; 10(1):88–98.
- LEE S. J.; HEYDECKER B.; KIM Y.H.; SHON E. Y. Dynamic OD estimation using three phase traffic flow theory. **Journal of Advanced Transportation**, 2011; 45:143–158.
- PERRAKIS K.; KARLIS D.; COOLS M.; JANSSENS D.; VANHOOF K.; WETS G. A Bayesian approach for modeling origin-destination matrices. **Transportation Research Part A: Policy and Practice**, vol. 46, no. 1, pp. 200–212, 2012.
- BUGEDA J. B.; MERCADÉ L. M.; MARQUÉS L.; CARMONA C. A Kalman-filter approach for dynamic OD estimation in corridors based on Bluetooth and Wi-Fi data collection. in **Proceedings of the 12<sup>th</sup> World Conference on Transportation Research**, 2010.



BARCELO J.; MONTERO L.; MARQUES L. Travel time forecasting and dynamic origin–destination estimation for freeways based on Bluetooth traffic monitoring. **Transportation Research Record**, 2010; 2175:19–27.

VAN DER ZIJPP N. Dynamic OD-matrix estimation from traffic counts and automated vehicle identification data. **Transportation Research Record**, 1997; 1607:1–18.

DIXON M. P.; RILETT L. R. Real-time OD estimation using automatic vehicle identification and traffic count data. **Journal of Computer Aided Civil Infrastructure Engineering**, 2002; 17(1):7–21.

FENG Y.; SUN J.; CHEN P. Vehicle trajectory reconstruction using automatic vehicle identification and traffic count data. **Journal of advanced transportation**, 2015, 49(2): 174–194.

ZHOU X. S.; MAHMASSANI H. S. Dynamic origin-destination demand estimation using automatic vehicle identification data. **IEEE Transactions on Intelligent Transportation Systems**, 2006; 17(1):105–114.

RAO W.; WU Y.; XIA J.; OU J.; KLUGER R. Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. **Transportation Research Part C: Emerging Technologies**, vol. 95, pp. 29–46, 2018.

CASTILLO E.; GALLEGO I.; MENÉNDEZ J. M.; RIVAS A. Optimal use of plate-scanning resources for route flow estimation in traffic networks. **IEEE Transactions on Intelligent Transportation Systems**, vol. 11, no. 2, pp. 380–391, 2010.

FU C.; ZHU N.; MA S. A stochastic program approach for path reconstruction oriented sensor location model. **Transportation Research Part B: Methodological**, vol. 102, pp. 210–237, 2017.

FEDEROV A.; NIKOLSKAIA K.; IVANOV S.; SHEPELEV V.; MINBALEEV A. Traffic flow estimation with data from a video surveillance camera. **Journal of Big Data**, vol. 6, no. 1, p. 73, 2019.

CASTILLO E.; MENÉNDEZ J. M.; JIMÉNEZ P. Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. **Transportation Research Part B**, 2008a; 42(5):455–481.

CASTILLO E.; MENÉNDEZ J. M.; CAMBRONERO S.S. Traffic estimation and optimal counting location without path enumeration using Bayesian networks. **Computer-Aided Civil and Infrastructure Engineering**, 2008b; 23(3):189–207.

- SÁNCHEZ-CAMBRONERO S.; ÁLVAREZ-BAZO F.; RIVASA.; GALLEGO I. A new model for locating plate recognition devices to minimize the impact of the uncertain knowledge of the routes on traffic estimation results. **Journal of Advanced Transportation**. 2020, 2020.
- YANG Y.; LIU J.; SHANG P.; XU X.; CHEN X. Dynamic Origin-Destination Matrix Estimation Based on Urban Rail Transit AFC Data: Deep Optimization Framework with Forward Passing and Backpropagation Techniques. **Journal of Advanced Transportation**. 2020, 2020
- ZHANG C.; WONG J.; LAI J.; YANG X.; SU Y., DONG Z. Extracting Origin-Destination with Vehicle Trajectory Data and Applying to Coordinated Ramp Meterings. **Journal of Advanced Transportation**, 2019, 2019. doi:10.1155/2019/8469316
- TEKNOMO K.; FERNANDE P. A theoretical foundation for the relationship between generalized origin-destination matrix and flow matrix based on ordinal graph trajectories. **Journal of Advanced Transportation**, 2012. doi:10.1002/atr.1214.
- PARRY K.; HAZELTON M. L. Estimation of origin-destination matrices from link counts and sporadic routing data **Transportation Research Part B**, 2012; 46(1):175–188.
- KWON J.; VARAIYA P. Real-time estimation of origin-destination matrices with partial trajectories from electronic toll collection tag data. **Transportation Research Record**, 2006; 1923:119–126.
- JAYAKRISHNAN R.; MAHMASSANI H. S.; HU T.Y. An evaluation tool for advanced traffic information and management systems in urban networks. **Transportation Research Part C** 1994; 3(2):129–147.
- ÁSMUNDSDÓTTIR R. Dynamic OD Matrix Estimation Using Floating Car Data **Delft University of Technology**, 2008.
- YANG X.; LU Y.; HAO W. Origin-destination estimation using probe vehicle trajectory and link counts. **Journal of Advanced Transportation**, vol. 2017, Article ID4341532, 18 pages, 2017.
- HUANG Z.; LING X.; WANG P. Modeling real-time human mobility based on mobile phone and transportation data fusion. **Transportation Research Part C: Emerging Technologies**, vol. 96, pp. 254–269, 2018.
- IBARRA-ESPINOSA S.; YNOUE R.; GIANNOTTI M.; ROPKINS K.; DE FREITAS E. D. Generating traffic flow and speed regional model data using internet GPS vehicle records. **MethodsX**, vol. 6, pp. 2065–2075, 2019.

MOREIRA-MATIAS L.; GAMA J., FERREIRA M.; MENDES-MOREIRA J.; DAMAS L. Time-evolving O-D matrix estimation using high-speed GPS data streams **Expert Systems with Applications**, vol. 44, pp. 275–288, 2016.

EISENMAN S. M. and LIST G. F. Using probe data to estimate OD matrices in Proceedings of the **7th International IEEE Conference on Intelligent Transportation Systems** (ITSC '04), pp. 291–296, Washington, DC, USA, October 2004.

HERRERA J. C.; BAYEN A. M. Traffic flow reconstruction using mobile sensors and loop data **The 87th Transportation Research Board Annual Meeting**. Washington D.C., 2008.

TOOLE J. L.; COLAK S.; STURT B.; ALEXANDER L. P.; EVSUKOFF A.; GONZÁLEZ M. C. The path most traveled: travel demand estimation using big data resources. **Transportation Research Part C: Emerging Technologies**, vol. 58, pp. 162–177, 2015.

ZIN T. A.; KYAING K.; LWIN K. K.; SEKIMOTO Y. Estimation of originating-destination trips in Yangon by using big data source. **Journal of Disaster Research**, vol. 13, no. 1, pp. 6–13, 2018.

YANG X.; ZOU Y.; TANG J.; LIANG J.; LJAZ M. Evaluation of short-term freeway speed prediction based on periodic analysis using statistical models and machine learning models. **Journal of Advanced Transportation**, vol. 2020, Article ID 9628957, 16 pages, 2020.

SÁNCHEZ-CAMBRONERO S.; CASTILLO E.; MENÉNDEZ J. M.; JIMÉNES P. Dealing with error recovery in traffic flow prediction using Bayesian networks based on license plate scanning data. **Journal of Transportation Engineering**, vol. 137, no. 9, pp. 615–629, 2010.

BAI J.; CHEN Y. A deep neural network based on classification of traffic volume for short-term forecasting. **Mathematical Problems in Engineering**, vol. 2019, Article ID 6318094, 10 pages, 2019.

LIU Z.; LI Z.; WU K.; LI M. Urban traffic prediction from mobility data using deep learning. **IEEE Network**, vol. 32, no. 4, pp. 40–46, 2018.

## Glossary

In this section, a list of frequently utilized terms, as to abbreviations related to the subject of this study. For the elaboration of the list, the study utilized the literature reviewed, as to publications from government agencies.

Terms are presented in alphabetical order.

**ALPR** (Automatic License Plate Recognition) - Automatic vehicle license plate recognition (also: ANPR, APR, ALR, ARPI, CLI, CLPR, CPR, LAP, LAPI, LPI, LPR, NPR).

**ALR** (Automatic License Recognition) - Automatic vehicle license plate recognition (also: ANPR, ANPR, ALPR, ARPI, CLI, CLPR, CPR, LAP, LAPI, LPI, LPR, NPR).

**ANPR** (Automatic Number Plate Recognition) - Automatic vehicle license plate recognition (also: APR, ALPR, ALR, ARPI, CLI, CLPR, CPR, LAP, LAPI, LPI, LPR, NPR).

**APR** (Automatic Plate Recognition) - Automatic vehicle license plate recognition (also: ANPR, ALPR, ALR, ARPI, CLI, CLPR, CPR, LAP, LAPI, LPI, LPR, NPR).

**ARPI** (Automatic Registration Plate Identification) - Automatic vehicle license plate recognition (also: ANPR, APR, ALPR, ALR, CLI, CLPR, CPR, LAP, LAPI, LPI, LPR, NPR).

**AVI** (Automatic Vehicle Identification) - Automatic identification of a vehicle, through different methods, such as OCR, ETC, etc.

**AVIS** (Automatic Vehicle Identification System) - A system of automatic vehicle identification.

**AVRS** (Automatic Vehicle Recognition System) - A system of automatic vehicle recognition.

**CLI** (Car License Identification) - Automatic recognition of a vehicle license plate (also: ANPR, APR, ALPR, ALR, ARPI, CLPR, CPR, LAP, LAPI, LPI, LPR, NPR).

**ETC** (Electronic Toll Collection) - Automatic recognition of a vehicle, for toll collection charging purposes.

**ITS** (Intelligent Transportation Systems or Intelligent Transport Systems) - Intelligent transport systems

**LPI** (License Plate Identification) - Automatic license plate identification. (also: ANPR, APR, ALPR, ALR, ARPI, CLI, CLPR, CPR, LAP, LPR, NPR).

**LPR** (License Plate Recognition) - Automatic license plate recognition. (also: ANPR, APR, ALPR, ALR, ARPI, CLI, CLPR, CPR, LAP, LPI, NPR).

**NPR** (Number Plate Recognition) - Automatic license plate recognition. (also: ANPR, APR, ALPR, ALR, ARPI, CLI, CLPR, CPR, LAP, LPI, LPR).

**License plate** - Consisting of codes, generally national, that represent the identification of the vehicle in each country.

**OCR** (Optical Character Recognition) - Process of scanning text images, with identification of the characters that make up the image.

**ODS** (Origin Demand Survey) - Surveys that aim to understand the trips in one roadway, usually about their origin and destination characteristics.

**SPS** (Stated Preference Survey) - Surveys that aim to obtain specific data to input in the modal choice models, such as interpretation of hypothetical alternatives during surveys.

**UE** (User Equilibrium) - User equilibrium traffic assignment model.

**VCS** (Vehicle count surveys) - Surveys that account for, in a determined period, the traffic that crosses a specific section.

## Appendix - Software and code

The software chosen for managing the database was SQL Server 2017, and the R programming language was applied to crunch the data. Later this dissertation moved into creating samples of the database in a plain text file to facilitate other contributors making use of the data.

The general query sequence was structured starting with database clean up and restructure to save disk memory, remove redundant fields, and move that information to accessory tables. Furtherly, routines of error checking, database sampling and simplification, and partial route identification were developed. Figure 59 shows the code leading to results of partial route reconstruction.

Figure 59: R code workflow

```
# Calculates partial route costs and merges with cost file for all starting
trip points adding N candidate t1 starting zones while also filtering zones
not passing through ETCs

dt <- dt0 %>% as.data.table() %>% setnames(.,c("orig","dest"),c("t2","t3"),
skip_absent = T) %>%

  .[as.data.table(cost_min), `:=`(cost23=cost), on=(t2=fid,t3=tid)] %>%
  .[as.data.table(cost_min), `:=`(cost32=cost), on=(t3=fid,t2=tid)] %>%
  .[is.na(cost23) | is.na(cost32), `:=`(cost23=dtm/60,cost32=dtm/60)] %>%
  .[t2==t3, `:=`(cost23=0,cost32=0)] %>%

merge(.,as.data.table(cost_min_f), by.x="t2", by.y="tid", allow.cartesian
=T) %>%

  setnames(., c("cost","fid","fidg",'fzclust',"atr_fid"),c("cost12","t1","t
1g","t1zc","atr_t1"),skip_absent = T) %>%

  .[, `:=`(tidg=NULL,dist=NULL,pass_avi=NULL,atr_tid=NULL,tzclust=NULL)] %>%
  .[t1>30000] %>%

# Calculates update trip costs and eliminates pairs with alternate, lower c
osting routes

  .[as.data.table(cost_min), `:=`(cost_min13=cost+5/60), on=(t1=fid,t3=tid
)] %>%
  .[as.data.table(cost_min), `:=`(cost_min12=cost+5/60), on=(t1=fid,t2=tid
)] %>%

  .[,cost13:=cost12+cost23] %>%
  .[(cost13<=cost_min13 & cost23>=5/60) | cost23<5/60] %>%
  .[,cost13:=cost13+5/60] %>%
  .[,atr_t1c:=atr_t1/(cost13^gfct)]
```

```

cols_atrk_t1 <- list("atr_t1c")
zones_f_t1 <- lapply(cols_atrk_t1, function(x,dt) {
  setorderv(dt, cols = x,order = -1) %>% .[, head(.SD, 50), by = c("t1g",
"t2","t3")]
  },dt=dt) %>% bind_rows(.) %>% as.data.table(.) %>% .[,c("t1","t2","t3"),w
ith=F] %>% unique(.)

# Calculates partial route costs and merges with cost file for all ending t
rip points adding N candidate t4 ending zones while also filtering zones no
t passing through ETCs

dt2 <- dt %>% semi_join(zones_f_t1,by=c('t1'='t1','t2'='t2','t3'='t3')) %>%
as.data.table(.) %>%

merge(.,as.data.table(cost_min_f), by.x="t3", by.y="fid", allow.cartesian
=T) %>%

setnames(., c("cost","tid","tidg",'tzclust',"atr_tid"),c("cost34","t4","t
4g","t4zc","atr_t4"),skip_absent = T) %>%

.[, `:=` (fidg=NULL,dist=NULL,pass_avi=NULL,atr_fid=NULL,fzclust=NULL)] %>%

.[t4>30000] %>%

# Calculates update trip costs and eliminates pairs with alternate, lower c
osting routes

.[as.data.table(cost_min), `:=` (cost_min24=cost+5/60), on=(t2=fid,t4=tid
)] %>%

.[as.data.table(cost_min), `:=` (cost_min34=cost+5/60), on=(t3=fid,t4=tid
)] %>%

.[, `:=` (cost14=cost12+cost23+cost34,cost24=cost23+cost34)] %>%

.[(cost24<=cost_min24 & cost13>=5/60) | cost13<5/60] %>%

.[, `:=` (cost14=cost14+5/60,cost24=cost24+5/60)] %>%

.[,atr_t14c:=atr_t1*atr_t4/(cost14^gfct)]

cols_atrk_t14 <- list("atr_t14c")
zones_f_t4 <- lapply(cols_atrk_t14, function(x,dt) {
  setorderv(dt, cols = x,order = -1) %>% .[, head(.SD, 100), by = c("t2","t
3","t4g")]
  },dt=dt2) %>% bind_rows(.) %>% as.data.table(.) %>% .[,c("t1","t2","t3","t4
"),with=F] %>% unique(.)

# Calculates update trip costs for t1 through t4 and eliminates pairs with
alternate, lower costing routes

dt3 <- dt2 %>% semi_join(zones_f_t4,by=c('t4'='t4','t2'='t2','t3'='t3')) %>
% as.data.table(.) %>%

.[as.data.table(cost_min), `:=` (cost_min14=cost+10/60), on=(t1=fid,t4=ti
d)] %>%

```

```

.[,exclude:=NA] %>% .[as.data.table(group_exceptions), `:=`(exclude=F), o
n=(t1g=group_orig,t4g=group_dest)] %>%

.[cost14<=cost_min14 | (exclude==F)]

# Distributes partial route flows, based on gravitational-based flow model
dt4 <- dt3 %>%

.[, `:=`(atr_t14c=atr_t1*atr_t4/(cost14^gfct))] %>%
setorderv(.,cols = "atr_t14c",order = -1) %>%

.[,(c("atr_t14c_g")):=lapply(.SD, function(x) {sum(x,na.rm=T)}),.SDcols=c
("atr_t14c"),by=c("t2","t3")] %>%

.[, `:=`(per_dist1=atr_t14c/atr_t14c_g)] %>%

.[,(c("per_dist1t")):=lapply(.SD, function(x) {sum(x,na.rm=T)}),.SDcols=c
("per_dist1"),by=c("t2","t3")] %>%

.[, `:=`(l1d`=`l1`*per_dist1)] %>%

.[,lapply(.SD, sum),.SDcols=c("l1d"),by=c("t1","t4")] %>%
setnames(., c("l1d"),c("l1"),skip_absent = T)

```

Source: Author

Figure 60 shows the detailed commented code for extrapolating and distributing partial routes

Figure 60: R code workflow

```

fdata <- initdata[,1:5] %>% .[as.data.table(equips_all), `:=`(id_new=id_new,
in_net=in_net,toll=ifelse(tipo=='avi',1,0)), on=(id)] %>%

.[id_new %in% points_graph$id_new] %>% setorder(.,vehicle_id,timestamp) %
>%

# compares next records vehicle license to establish a trip end flag

.[, `:=`(diff_veic_lead=shift(vehicle_id,1,type = 'lead')!=vehicle_id)] %>
%

# write inline next locations and timestamps

.[, `:=`(id_next=shift(id,1,type = 'lead'),id_new_next=shift(id_new,1,type
= 'lead')
,timestamp_next=shift(timestamp,1,type = 'lead'))] %>%

# for the next record being from a different vehicle all time and locatio
ns columns are set to NA

.[diff_veic_lead==T, `:=`(id_next=NA,id_new_next=NA,timestamp_next=NA)] %>
%

# calculates the duration in hours of the time passed from current until
next location

```



```

.[diff_veic_lead==F, hours23:=round(as.numeric(difftime(timestamp_next, timestamp, units = "hours")),3)] %>%
  # removes movements that take 0h (duplicate database identifications)
.[hours23>0 | is.na(hours23)] %>%
  # recreate different vehicle flag after duplicate removal
.[, `:=`(diff_veic_lead=shift(vehicle_id,1,type = 'lead')!=vehicle_id)] %>%
%
# rewrite inline next locations and timestamps
.[, `:=`(id_next=shift(id,1,type = 'lead')
  ,id_new_next=shift(id_new,1,type = 'lead')
  ,timestamp_next=shift(timestamp,1,type = 'lead'))] %>%
  # for the next record being from a different vehicle all time and locations columns are set to NA
.[diff_veic_lead==T, `:=`(id_next=NA, id_new_next=NA, timestamp_next=NA)] %>%
%
# compares previous records vehicle license to establish a trip start flag
.[, `:=`(diff_veic_lag=shift(vehicle_id,1,type = 'lag')!=vehicle_id)] %>%
  # write inline previous locations and timestamps
.[, `:=`(id_prev=shift(id,1,type = 'lag'), id_new_prev=shift(id_new,1,type = 'lag')
  ,timestamp_prev=shift(timestamp,1,type = 'lag'))] %>%
  # for the previous record being from a different vehicle all time and locations columns are set to NA
.[diff_veic_lag==T, `:=`(id_prev=NA, id_new_prev=NA, timestamp_prev=NA)] %>%
  # recalculate new movement times
.[diff_veic_lead==F, hours23:=round(as.numeric(difftime(timestamp_next, timestamp, units = "hours")),3)] %>%
  # calculate prior movement times
.[diff_veic_lag==F, hours12:=round(as.numeric(difftime(timestamp, timestamp_prev, units = "hours")),3)] %>%
  # merge minimum estimated time between current and next locations
.[as.data.table(cost_min[,c("fid", "tid", "cost")]), hours_min:=cost, on=(id_new=fid, id_new_next=tid)] %>%
  # creates a trip end flag on the basis of being different vehicles, time between current and next location over an estimated minimum and added 6 hours
.[, trip_break:=case_when(
  diff_veic_lead ~ 1,
  is.na(hours_min) ~ if_else(hours23>t,1,0),
  TRUE ~ if_else(hours23>(hours_min+t),1,0))] %>%
  # trip id creation with a cumulative sum of breaks

```

```

.[,trip_id:=cumsum(shift(trip_break,1,type = 'lag', fill = 0))] %>%
# flags identifying next and previous trips being different
.[,`:=`(diff_trip_lead=shift(trip_id,1,type = 'lead')!=trip_id,diff_trip_lag=shift(trip_id,1,type = 'lag')!=trip_id)] %>%
# in case of different trip for the next record, locations are set to NA
.[diff_trip_lead==T,`:=`(id_next=NA,id_new_next=NA)] %>%
# in case of different trip for the previous record, locations are set to NA
.[diff_trip_lag==T,`:=`(id_prev=NA,id_new_prev=NA)]

# memory dump
rm(initdata);gc()

# determine every unique combination of previous, current and next locations
# for the database and calculates the distance between each of them
pairs <- rbind(fdata[,c('id_prev','id'),with=F] %>% na.omit(.) %>% unique(.)
) %>% setnames(c('id1','id2'))
          ,fdata[,c('id','id_next'),with=F] %>% na.omit(.) %>% unique(.)
) %>% setnames(c('id1','id2'))
          ,fdata[,c('id_prev','id_next'),with=F] %>% na.omit(.) %>% unique(.)
) %>% setnames(c('id1','id2')) %>%
unique(.) %>%
.[as.data.table(equips_all),`:=`(x1=x,y1=y), on=.(id1=id)] %>%
.[as.data.table(equips_all),`:=`(x2=x,y2=y), on=.(id2=id)] %>%
.[,dist:=distCosine(cbind(x1,y1),cbind(x2,y2))/1000]

fdata <- fdata %>%
# merges the pairs table to the database (faster than calculating every pairs distance)
.[pairs, `:=`(dist12=dist), on=.(id_prev=id1,id=id2)] %>%
.[pairs, `:=`(dist13=dist), on=.(id_prev=id1,id_next=id2)] %>%
.[pairs, `:=`(dist23=dist), on=.(id=id1,id_next=id2)] %>%
# determine estimated speed of the displacement (disregarding the path geometry)
.[,speed12:=if.na(round(dist12/hours12,0))] %>%
.[,speed23:=if.na(round(dist23/hours23,0))] %>%
# removes trips with speed higher than defined speed of 120kph
.[!trip_id %in% unique(fdata[speed23>speed_max | speed12>speed_max]$trip_id)] %>%
# calculates angles and creates a trip break flag on angles under 30 degrees

```

```

.[(dist12*dist23)>0,angle_turn:=360/(2*pi)*acos((dist12^2+dist23^2-dist13
^2)/(2*dist12*dist23))] %>%

.[(dist12*dist23)==0,angle_turn:=0] %>%

.[,trip_break:=trip_break+ifelse(diff_trip_lead,0,if_else(shift(angle_tur
n,1,type = 'lead',fill = 0)<angle,1,0,0))] %>%

# recreates trip ids based on newly established break points

.[,trip_id:=cumsum(shift(trip_break,1,type = 'lag',fill = 0))] %>%

# in case of different trip for the next record, locations are set to NA

.[,diff_trip_lead:=shift(trip_id,1,type = 'lead',fill = T)!=trip_id] %>%

.[,id_new_next:=ifelse(diff_trip_lead,NA,shift(id_new,1,type = 'lead'))]
%>%

# remove movements not captured on network, whenever the current or next
location is the same (due to simplification) the record is removed

.[id_new!=ifelse(is.na(id_new_next),0,id_new_next)] %>%

# trip sequence of movements

.[,`:=`(seq=1:.N,first=first(id_new),last=last(id_new)),by="trip_id"] %>%

# trip sequence simplified to only represent first and last radar equipme
nt and avi's in between

.[seq==1 | seq>=shift(seq,1,type='lead') | id_new %in% filter(equips_all,
tipo=='avi')$id_new] %>%

.[,`:=`(trip_seq=paste(id_new, collapse = ","),by="trip_id"]

# filters trips in network

fdata_fil <- fdata %>% filter(in_net==1)

# determines the amount of trips for each simplified trip sequence, first a
nd last equipment and vehicle category

trips <- unique(fdata_fil[,c('trip_id','trip_seq','first','last','cat','tol
l')]) %>%

.[,.(vehicle_count=.N,toll=max(toll)),by=list(trip_seq,first,last,cat)]

# determines the amount of trips a municipality generates based on vehicles
with registered license plates

trips_city_count <- unique(fdata[,c('trip_id','city_id','cat','toll')]) %>%

.[!is.na(city_id),.(vehicle_count=.N),by=list(city_id,cat,toll)]

```

Source: Author