FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Modelling Occupancy Estimation in Public Transport: A Data-Driven Approach

Margarida Ribeiro Cosme

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Teresa Galvão Dias

Co-Supervisor: Prof. Marta Campos Ferreira

July 28, 2023

Modelling Occupancy Estimation in Public Transport: A Data-Driven Approach

Margarida Ribeiro Cosme

Mestrado em Engenharia Informática e Computação

July 28, 2023

Resumo

Esta dissertação explora o potencial da estimativa de ocupação em transportes públicos, com base em dados de bilhética da Área Metropolitana do Porto, recolhidos pelos sistemas de Recolha Automatizada de Tarifas (AFC). Os dados utilizados para este fim consistem em informações sobre o embarque de passageiros.

Essas informações são úteis para entender a procura dos transportes públicos e auxiliar na tomada de decisões informadas sobre o planeamento de transporte, por exemplo, na capacidade dos veículos e no planeamento de frequência de viagens. Esta pesquisa também pode ajudar os utilizadores de transportes público a planear melhor os seus trajetos e, potencialmente, na toma de decisões relacionadas com a lotação do autocarro, como por exemplo esperar por um autocarro vazio mais tarde.

O estudo começa com a realização de uma rigorosa revisão da literatura, que engloba uma análise abrangente das metodologias existentes e do estado da arte, predominante na estimativa da ocupação.

O modelo é desenvolvido usando uma abordagem data-driven, envolvendo um rigoroso processo de limpeza de dados, e uma análise completa dos dados dos sistemas AFC para extrair conhecimento valioso, sobre padrões, tendências e outras relações existentes na informação das viagens e dos passageiros. Informado pelos insights obtidos, o modelo de estimativa de ocupação é desenvolvido e, posteriormente, submetido a um processo de validação, para avaliar a sua precisão. Isso é feito com a ajuda de um conjunto de dados auxiliar, que contêm informações sobre as saídas dos passageiros.

O modelo demonstra um bom desempenho com baixos valores de erro para o contexto em causa, e uma alta correlação entre os valores estimados e os reais, fornecendo uma ferramenta confiável para estimativa da ocupação do veículo.

Além do modelo bem-sucedido, o trabalho revelou alguns insights valiosos sobre padrões de viagem de passageiros, contribuindo para o corpo de conhecimento em Sistemas de Transporte Inteligentes e estimativa de fluxo de passageiros usando dados de sistemas AFC.

Trabalho futuro inclui a incorporação de fatores adicionais, como do tipo de bilhete no modelo, e inclui o aproveitamento do conjunto de dados auxiliares, para propósitos de feature engineering e de calibração do modelo. A exploração de técnicas de machine learning para aprimorar ainda mais a precisão do modelo também é incentivada.

Abstract

This dissertation explores the potential for occupancy estimation in public transport, based on Metropolitan Area of Porto's ticketing data, collected by Automated Fare Collection (AFC) systems. The data used for this purpose consists in passenger boardings information.

This information is useful for understanding public transports' demand and assisting informed decision making regarding transport planning, for instance vehicle capacity and trips frequency planning. This research can also assist public transport users to better plan their commutes and potentially make decisions related to the bus occupancy, such as waiting for a later, empty bus.

The study begins with the conduction of a rigorous literature review, which encompasses a comprehensive examination of existing methodologies and the prevailing state of the art in occupancy estimation.

The model is developed using data-driven approach, involving a rigorous data cleaning process, and a thorough analysis of the AFC data to extract valuable knowledge, regarding passenger travel patterns, trends, and other existing relationships in the data. Informed by the gained insights, the occupancy estimation model is developed, and subsequently put through a validation process in order to access its accuracy. This is done with the help of an auxiliary dataset, containing passengers' exits information.

The model demonstrates a good performance with low error values for the context, and a high correlation between the predicted values and the actual ones, providing a reliable tool for vehicle occupancy estimation.

In addition to the successful model, the work revealed some valuable insights regarding passenger travel patterns, contributing to the body of knowledge in Intelligent Transportation Systems and passenger flow estimation using AFC systems' data.

Future work includes incorporating additional factors such as ticket type into the model, leveraging the auxiliary dataset for feature engineering and model calibration, and exploring machine learning techniques to further enhance predictive capabilities.

Acknowledgments

Firstly, I would like to express my deepest gratitude to my supervisor, professor Teresa Galvão Dias, and my co-supervisor, professor Marta Campos Ferreira, for their invaluable guidance and support throughout this project. Their expertise, insights, and encouragement have been instrumental in shaping the outcome of this work, and contributed to the growth of my skills and knowledge.

My sincere appreciation goes to my parents and sister for the constant support, understanding and motivation, during the entirety of my studies, and throughout this endeavor in particular. Their love, encouragement, and belief in me have been my driving force. I am grateful to my feline companion, my cat Joy, that kept me company for most of the writing process, providing love, comfort, and always reminding me to take breaks.

I wish to express my heartfelt appreciation to all my friends that accompanied me during my studies, for all the memorable times, and for the motivation and endless study sessions, in particular during the duration of this project.

I am deeply grateful to TIP and to STCP for providing the necessary resources and facilities that have enabled the successful completion of this thesis. Their commitment to excellence has been instrumental in the realization of this endeavor.

I am indebted to the numerous researchers and scholars whose prior work and contributions have paved the way for this study. Their groundbreaking research and discoveries have been a source of inspiration and knowledge.

I would also like to acknowledge ChatGPT, that was consistently used as an assistance tool during the development of this project, mostly for code development and debugging, but also for brainstorming and document structuring.

Lastly, I would like to express my deepest gratitude to anyone else who has directly or indirectly contributed to this thesis, I would not have been able to finish this chapter without all of the people who helped me along the way.

Margarida Cosme

"A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

Herbert A. Simon

Contents

1 Introduction **1**
1.1 Problem Context . . . 1
1.2 Motivation . . . 1
1.3 Research Objectives . . . 2
1.4 Dissertation Structure . . . 2

2 Passenger Occupancy Estimation in Public Transport **3**
2.1 WLAN Probing . . . 3
2.2 Sensor-Supported Systems . . . 4
2.2.1 Radar Sensing . . . 4
2.2.2 Infrared Sensing . . . 5
2.2.3 Weight and Pressure Sensing . . . 5
2.2.4 Vision-Based Systems . . . 6
2.3 Automatic Fare Collection Systems . . . 8
2.3.1 Origin Destination Estimation . . . 9
2.3.2 Passenger Loads Estimation . . . 12
2.3.3 Limitations . . . 17

3 Problem Characterization **18**
3.1 Domain . . . 18
3.1.1 Metropolitan Area of Porto . . . 18
3.1.2 Andante System . . . 19
3.1.3 Anda App . . . 19
3.1.4 AFC Data . . . 19
3.2 PART: 2019 Context . . . 20
3.3 Case Study: Line 205 . . . 21
3.3.1 Stops . . . 21
3.3.2 Variants . . . 23

4 Methodology **25**
4.1 Technologies . . . 25
4.2 Data Description . . . 25
4.2.1 Andante Dataset . . . 25
4.2.2 Anda Dataset . . . 26
4.2.3 Auxiliary Datasets . . . 27
4.3 Data Preparation . . . 29
4.3.1 Data Selection . . . 29
4.3.2 Data Integration . . . 29

4.3.3 Data Cleaning . . . 30
4.3.4 Data Transformation . . . 34
4.4 Line Treatment . . . 34
4.4.1 Mitigating Illogical Values . . . 35
4.4.2 Grouping Data by Trip . . . 37
4.4.3 Trip Concept Definition . . . 38
4.5 Anda Treatment . . . 39
4.6 Data Analysis . . . 40
4.6.1 Daily Patterns Analysis . . . 40
4.6.2 Stops Patterns Analysis . . . 44
4.6.3 Entries and Exits Patterns Analysis . . . 49
4.7 Bus Line Occupancy Estimation Model . . . 52
4.7.1 Assumptions and Insights from Data Analysis . . . 54
4.7.2 Occupancy Estimation Process . . . 55
4.7.3 Methodology Pipeline . . . 57

5 Results and Discussion 58
5.1 Evaluation Method . . . 58
5.1.1 Data Selection . . . 58
5.1.2 Model Evaluation . . . 58
5.1.3 Evaluation Metrics . . . 59
5.2 Results . . . 60
5.2.1 Outward vs Return Trips . . . 64
5.2.2 Line Variant . . . 65
5.2.3 Working Day vs Weekend Trips . . . 66
5.2.4 Daily Period . . . 67
5.2.5 Seasonality . . . 68
5.2.6 Discussion . . . 69
5.3 Limitations . . . 69
5.3.1 Model Limitations . . . 69
5.3.2 Validation Limitations . . . 70

6 Conclusions and Future Work 71
6.1 Conclusions . . . 71
6.2 Future Work . . . 72

References 73

List of Figures

3.1 Zoning System 2017 20
3.2 Zoning System 2019 20
3.3 Line 205 Variants 23
3.4 Line 205 Vehicle routes in a weekday 24

4.1 Total Demand by Month and day of Week 40
4.2 Andante Demand by Hour and day of Week 41
4.3 Andante Demand by Hour Weekend 41
4.4 Direction 1: Demand by Stop and Hour 42
4.5 Direction 2: Demand by Stop and Hour 43
4.6 Andante: First and Last Daily Validation Demand by Hour 43
4.7 Andante: Total Demand Distribution by Stop and Direction 44
4.8 CMP-CQ: Demand Distribution by Stop and Direction 45
4.9 SR-CQ: Demand Distribution by Stop and Direction 46
4.10 SR-LDD: Demand Distribution by Stop and Direction 46
4.11 Anda: Entry Zones Direction 1 and Exit Zones Direction 2 49
4.12 Andante: Entry Zones Direction 1 and Estimated Exit Zones Direction 2 49
4.13 Anda: Entry Zones Direction 2 and Exit Zones Direction 1 50
4.14 Andante: Entry Zones Direction 2 and Estimated Exit Zones Direction 2 50
4.15 Anda: Direction 1 Entries and Direction 2 Exits by Stop 51
4.16 Anda: Direction 2 Entries and Direction 1 Exits by Stop 51
4.17 Methodology Pipeline 57

5.1 Occupation MAE, NMAE, RMSE, NRMSE and ESV Box plots 61
5.2 Worst performing trip: Actual and Predicted Occupancy 62
5.3 Best performing trip: Actual and Predicted Occupancy 62
5.4 Trip with the most passengers: Actual and Predicted Occupancy 63
5.5 Trip with the least passengers: Actual and Predicted Occupancy 64

List of Tables

2.1 Summary of previous research in OD estimation using AFC systems 13

3.1 Line 205 route 22
3.2 Line 205 Variants 23

4.1 ROD and RR Attributes description and characteristics 27
4.2 Anda Attributes description 28
4.3 Data Integration's impact on the datasets 30
4.4 Outliers identification criteria and results 32
4.5 Data Cleaning results by phase 33
4.6 Variants Identification in Andante 39
4.7 Percentage difference between Andante and Anda Validation distributions by stop 48
4.8 Anda: Percentage difference between opposite direction's entries and exits 53

5.1 Overall Occupation Estimation Results 60
5.2 Overall Exits Estimation Results 63
5.3 Estimation Results by Direction 64
5.4 Estimation Results by Variant 65
5.5 Estimation Results by Weekday 66
5.6 Estimation Results by Period 67
5.7 Estimation Results by Season 68

Abbreviations and Symbols

AFC	Automated Fare Collection
APC	Automated Passenger Counter
APC-VL	Automated Passenger Counter with Vehicle Location
AVL	Automatic Vehicle Location
EVS	Explained Variance Score
FEUP	Faculty of Engineering, University of Porto
GPS	Global Positioning System
GTFS	General Transit Feed Specification
ITS	Intelligent Transport Systems
MAE	Mean Absolute Error
MAP	Metropolitan Area of Porto
NMAE	Normalized Mean Absolute Error
NRMS	Normalized Root Mean Square Error
OD	Origin-Destination
PIR	Passive Infrared
RMSE	Root Mean Square Error
RF	Radio Frequency
STCP	Sociedade de Transportes Colectivos do Porto
TCM	Trip-Chaining Method
TIP	Transportes Intermodais do Porto
WIM	Weight-In Motion
WLAN	Wireless Local Area Network

Chapter 1

Introduction

1.1 Problem Context

This dissertation aims to investigate the potential for occupancy estimation in public transport vehicles based on Automated Fare Collection (AFC) system data. The research focuses on the Porto Metropolitan area and aims to develop a model that can accurately estimate the occupancy of public transport vehicles based on ticket validation data. The work begins with a comprehensive literature review of existing methods of occupancy estimation and the current state of the art. The methodology involves analyzing the AFC data of the Porto metropolitan area, developing a model for occupancy estimation, and finally validating the model to access its accuracy. This chapter describes the research motivation, the problem and research objectives, and, finally, the dissertation structure.

1.2 Motivation

The development of a public transport vehicle occupancy estimation model is relevant in many ways as it touches upon various important aspects of urban organization, smart cities intelligent transport systems, sustainable development, and fleet management.

With the increasing population and increasing urbanization, the demand for efficient public transport is increasing, making it crucial to have a good public transport system that is organized, efficient and meets the demand. The optimization of public transport is a key factor in promoting its use, contributing towards a sustainable development, both environmentally and socially, by decreasing the amount of private transportation, decreasing the amount of empty buses, and providing the population with a quality transportation option.[17]

One of the challenges of organizing a good public transport system is to have adequate supply, and the organization of operations. By knowing the occupancy of public transport vehicles in real-time, decision-makers can better understand the demand and plan the capacity and frequency of vehicles accordingly, optimizing their fleets, and reducing costs. [17]

Additionally, this model can help to address the challenges of urban organization, making it easier to plan and manage cities in a way that benefits all stakeholders, including residents, businesses and visitors.

1.3 Research Objectives

- To analyze and review the existing literature on the extraction of knowledge from AFC data and vehicle occupancy estimation models.
- To conduct an analysis of the AFC data from the Metropolitan Area of Porto (MAP) public transport to understand the travel patterns and validations carried out by passengers.
- To develop a public transport vehicle occupancy estimation model based on the AFC data from ticket validations.
- To validate the developed model and evaluate its performance.
- To determine the feasibility and usefulness of the model in assisting decision making regarding transport planning and helping passengers to plan their commutes.
- To contribute to the body of knowledge on vehicle occupancy estimation, particularly with regards to the use of Automated Fare Collection systems in public transport.

1.4 Dissertation Structure

In addition to the introduction, this thesis is organized in 5 more chapters. Chapter 2 consist in the establishment of the current state of art of passenger occupancy estimation methods in public transport. Chapter 3 describes the particular case study, while Chapter 4 is a comprehensive overview of the methodology followed. The results obtained are presented and discusses in Chapter 5, and finally, in Chapter 6 the conclusions and future work are presented.

Chapter 2

Passenger Occupancy Estimation in Public Transport

Human occupancy detection (HOD) is the process of detecting the presence of people in a given space. It has various applications, including transportation safety and efficiency, building automation, and security systems. [51] In the context of transportation, occupancy detection is used to estimate the number of passengers in autonomous or public vehicles, or at public transport stations. Many different methods have been proposed for occupancy estimation systems, Kuchár et al. (2023) categorise these techniques as invasive, when the technology employed for passenger counting is within the vehicle, and noninvasive, if it is in the exterior. Invasive methods can be implemented using various vision-based systems, mechanical systems, Wireless Local Area Network (WLAN) probing, etc. they provide high accuracy but can be expensive and require installation and maintenance. Non invasive methods occur most regularly through the vehicle windows, using vision-based systems. The variety of vehicle types present the biggest challenge because occupancy estimation methods, namely the placement of the sensors, need to be adjusted for each situation. [42]

2.1 WLAN Probing

With the proliferation of wireless technologies, occupancy estimation using WLAN probing, has gained significant attention in recent years. This method involves analyzing the WLAN signals transmitted by passengers' devices to determine the amount of people in a specific area.

Mikkelsen et al. (2016) implemented WLAN-based occupancy estimation using a simple algorithm in Aalborg, Denmark. The results showed that the proposed method could accurately estimate the occupancy of the bus in real-time, with the correct fine-tuning; the algorithm was very sensitive towards the algorithm parameter settings [56]. Building on this work Mehmood et al. (2019) used a similar method in Sydney Australia with similar results [55]. Vieira et al. (2020) proposed a mathematical model that considers the signal strength of WLAN probes, to estimate the occupancy of public transportation using WLAN probing in Belo Horizonte, Brazil.

The results showed that more complex algorithms such as mathematical models can improve the accuracy of the passengers' estimation. [80]

The use of WLAN-based occupancy estimation has several benefits, including non-intrusiveness, cost-effectiveness, and the ability to provide real-time data. However, there are also several flaws with this approach. The accuracy of WLAN-based occupancy estimation is affected by factors such as signal interference, passenger density [56], and the quantity and positioning of the WLAN sensors, that must to be optimized in order to mitigate the impact of devices outside the bus [55]. Furthermore, not all passengers carry devices with enabled Wi-Fi, and some passengers carry more than one device. [42]

2.2 Sensor-Supported Systems

In the context of passenger occupancy estimation in public transports, sensor-supported systems play a critical role in detecting human presence. These systems are essential for various applications, including controlling safety belts, airbags, warning devices, and logistics in mass transport systems including as buses, subways, trams, and railways. [76] The sensors used for human occupancy detection (HOD) can be classified into active sensing and passive sensing categories.

Active sensing modalities include lidar, radar, and ultrasonic sensors, which use a transmitter to send out a signal that is reflected off the target and then collected by a receiver. While active sensing can provide accurate and real-time data, these systems are expensive and can be obstructed or jammed. In contrast, passive sensing modalities such as weight and pressure sensing, infrared sensing, and vision-based systems rely on detecting or responding to certain environmental stimuli without emitting any signal. Passive sensing is generally considered non-invasive and low-priced, making it an attractive option for HOD systems. [51]

Different radar-based systems have been proposed for seat occupancy estimation, passive infrared (PIR) sensors have been used for the same purpose. Weight-in-motion (WIM) and pressure sensing systems can be used to determine the number of passengers on a bus, as well as vision-based systems, that use cameras to identify and track passengers. Each of these sensing modalities has its strengths and limitations, and their selection depends on the specific application and system requirements.

2.2.1 Radar Sensing

Radar sensors have been proposed as an alternative solution to traditional occupancy sensors due to their ability to detect human presence accurately, regardless of whether the human subject is moving at varying rates of activity or at rest. [42]

The quadrature Doppler radar, which monitors cardiac and breathing signals to enhance the detection of stationary subjects, has been proposed in Yavari et al. (2013). [86] Similarly, Lazaro et al. (2021) proposed a affordable pulsed mm-Wave radar for measuring seat occupation and breathing rate estimation, which does not require high complexity or computational costs. [45]

Furthermore, Sterner et al. (2012) used an Radio Frequency (RF) transmission system for passenger counting, which can detect weather a seat is occupied by analyzing the electromagnetic wave relationship between the sent and received waves. This system can be used for occupancy estimation in buses and other public transportation by installing the sensor system in the seats. [76]

Liu et al. (2020) proposed a cognitive radio human occupancy detection over RF analysis (CRhodora) system, which, utilising software-defined radio, scans the spectrum from the lowest to the highest frequency while inside a closed environment. Principal component analysis and recursive feature elimination algorithms are used to choose the frequency bands susceptible to human occupancy detection, and labels are assigned in accordance with the status of human occupancy. The trained classifier is then used for real-time occupancy detection. The proposed system offers several advantages such as the absence of invasive sensors, reconfigurable cognitive radio, and data efficiency. [51]

However, it is important to note that most radar-based radar-based occupancy detection research focuses on passenger detection in automobiles.

2.2.2 Infrared Sensing

Among the various sensors systems, PIR sensors are the most frequently used for detecting a person's presence, due to being widely available and due to their low cost. However, PIR sensors are primarily motion detectors and may lead to high rates of false alarms and inaccurate occupancy estimation. Additionally, PIR sensors cannot identify stationary objects, further limiting their effectiveness as a presence sensor. [86, 82]

However, they still have potential for occupancy estimation. Rahmatulloh et al. (2020) utilizes Internet of Things and PIR sensors for bus passenger monitoring. The PIR sensor detects passengers based on the degree of movement at which a person's body radiates heat. The PIR sensor is placed above passenger seats to prevent false detections, however, the detection accuracy of sensors is affected by passenger movement. Compared to infrared transmitter-receiver or switch sensors, the PIR sensor provides superior object detection precision, reaching 100% in 10 sensor tests. [71]

2.2.3 Weight and Pressure Sensing

Weight-in-motion (WIM) systems consist in existing traffic infrastructure that plays a critical role in the preservation of road infrastructure by providing real-time monitoring of traffic, vehicle loads, and condition-based maintenance. [23] Loga et al. (2016) used WIM systems, in concrete the measure of the total weight of the vehicles, for the estimation of vehicle occupancy rate in public transit. This technology's application for passenger estimation is constrained by the WIM station's accuracy, the vehicle's kerb weight and fuel level, as well as rough estimates of the passengers' weight that can vary significantly. [52]

Using a different approach, Zhu et al. (2008) developed a method for distinguishing the direction of passengers based on the pressure exerted when a person ascends or descends the stairs in the public transport. The authors used a pressure sensor pad to measure the pressure distribution caused by the movement of passengers on and off the bus. They then analyzed the data to determine the number of passengers occupying the bus. This method showed accuracy as high as 90% in passenger counting and above 80% in direction distinguishing, but it requires additional equipment and installation costs. [91]

2.2.4 Vision-Based Systems

Vision-based systems have been extensively applied to transports, particularly to autonomous vehicles. These technologies have been used in seat-belt detection, High-Occupancy Vehicle (HOV) lane surveillance, and airbag systems improvement. Several researchers have proposed stereo-based systems for intelligent airbag systems with the main focus being on seat classification and position estimation, particularly head tracking. Although some of the proposed systems have shown impressive results, the computation time remains a challenge. [42]

In public transportation, various vision-based systems have been proposed for passenger occupancy estimation, mostly non-invasively, by counting passengers at the entrance of a vehicle, including a bus [53, 49, 84, 19, 50, 20], train or metro [15], or at a station. [27]

Deparis et al (1996) and Khoudour et al. (1996) introduced a method using two active linear cameras to count passengers moving in the camera's field of view. [22, 38] Gerland et al (1999) proposed a system using one infrared sensor installed above the door frame to count and distinguish passengers. [27] Yang et al. (2010) proposed a framework for passenger counting based on feature-point tracking and clustering, which claims to perform better than approaches based on background modelling and foreground blob tracking. [84]

Amin et al. (2008) presented a technique for estimating individuals employing affordable, low-resolution visual and infrared cameras, including the IRISYS IRI1001 thermal imager. The system can accurately count people within a 3% error rate across various lighting conditions. [6]

To overcome the challenges of occlusion and changing lighting, Chen et al. (2008), and other research deployed zenithal cameras. [19, 20, 21, 49]

Chen et al. (2008) recorded the bilateral movement of passengers inside the bus using a zenithal camera. Each of the several blocks that make up the collected frame is categorized in accordance with its motion vector. Those blocks are then assumed to be part of the same motion object, in the case that the number of similar motion vector blocks exceeds a certain threshold. The proposed system can resolve the innate issues with the bus's lighting changes and camera instability obtaining an accuracy of 92%. [19]

Later, Chen et al. (2012) applied a method for detecting passenger heads using consistent hair color and solidity filling in head circles, through the application of the Canny algorithm as a pre-processing step to detect edges in the image before applying the Hough transform. The combination of these techniques helps overcome challenges such as cluttered background, lighting impact, and flexible body motions in the videos. [20] Later, the same authors employ Haar features and Adaboost algorithm for human head detection using OpenCV. [21]

Li et al. (2016) managed to accurately count the amount of passengers boarding and alighting a bus through RGB-D video and a zenithal camera mounted in the bus door. The method includes detection, tracking, and counting algorithms, and it is stable, accurate, and meets real-time requirements. [49]

Lengvenis et al. (2013) employed four methods to estimate public transport occupation, [48] while Hussain et al. (2020) highlighted early works based on simple video cameras. [33] Van et al. (2011) a method based on shapes and range data. [65]

Additionally, Perng et al. (2016) suggested an approach based on video processing, extracting moving objects, and tracking individuals to count them as they cross a counting line, through background subtraction. The system can adjust to varying lighting conditions and does not require any special background or physical lines on the ground. [67]

Qian et al. (2017) used a similar technique, [70] while Mukherjee et al. (2011) proposed a different approach for human detection, tracking, and validation in order to estimate crowding at the train station.[57] Belloc et al. (2018) presented a collection of data of people boarding and alighting a public transport vehicle, captured in a controlled laboratory environment that simulated a metro train carriage, providing baseline results for the future. [15]

Most of the previous mentioned research employed visible lights cameras and computer vision methods, however, there are other vision-based methods, such as time-of-flight (TOF) or stereo imaging, that require additional hardware.

Klauser et al. (2015) employed infrared sensors, to estimate occupancy, based on a time-of-flight system, developed using image processing, as well as statistical and probabilistic methods. However, this strategy proves to be effective exclusively if all objects can be differentiated from each other within at least one frame. [40]

Stereo vision systems have been extensively used for passenger counting in buses, and are unobtrusive, cost-effective and easily integrated. [16]

Yahiaoui et al. (2010) implemented a stereo vision system to extract three-dimensional information. To reconstruct the trajectories of the passengers boarding or alighting the bus during the image sequence, the marks describing their heads are tracked, then, passenger counting occurs based on those trajectories. This method was validated using numerous realistic experiments, including two large image sequences databases depicting realistic scenarios, where it obtained an accuracy of 97% and 99%. [83]

Bernini et al (2014) used a stereo vision system, to count passengers getting in or out of a bus, through feature-based methods. In this approach, a stereo camera is placed over the door of the bus, in a zenith position, to reduce the overlapping problem. [16]

Invasive passenger detection using imaging processing techniques has also been proposed by some studies recently. [62, 61, 60, 88] Nakashima et al. (2019) and Zhang et al. (2020) have employed deep learning techniques, in particular the YOLO network model. [60, 88]

Overall, vision-based approaches offer various techniques for passenger occupancy estimation, showing promising results in different scenarios. However, these methods are sensitive to noise and artefacts in the data, which can be caused by various factors such as saturation, white balance, and lens flare, as well as weather and light conditions. [41] Other limitations are that the accuracy typically decreases as the number of people in the scene increases, [16, 6]. Additionally the methods usually present additional costs in equipment and installation, [60] and can present passenger privacy concerns. [41]

2.3 Automatic Fare Collection Systems

Automatic Fare Collection (AFC) systems are a key component in the majority of Intelligent Transportation Systems (ITS), and consist in the use of smart travel cards as an alternative to the traditional tickets, improving transport systems' information flow, and providing a convenient and instant payment process to their users. [64] While the primary objective of AFC systems is the automation and efficient management of the fare collection ticketing process, the vast amount of detailed data gathered about onboarding or alighting transactions presents great opportunities to extract valuable knowledge about public transport passengers' behavior, useful for operators, transit planners and researchers. [81, 89]

Due to the complex nature of real-world scenarios, investigations in the transport industries require a significant amount of comprehensive data with a high level of precision and detail. Since the 1960s, collecting these datasets has been one of the most significant problems in the transit fields. [25] Some of the applications of this data consist in travel pattern mining [54, 18], origin-destination (OD) estimation, route choice modeling [39, 90], trip purpose inference [47, 44, 2], transit performance indicators, identification of policy alternatives, advertising, target marketing, and service planning. All stemming from the study of the demand and travel behavior of passengers.

Prior to AFC systems, the data required for transportation studies was commonly acquired through various expensive and time-consuming travel surveys, generating small sample sizes subject to response bias. [13]. Even though these methods generated datasets richer in the attributes compared to the AFC data, the missing properties from the AFC datasets can often be determined using other information sources, among them the road network, traffic conditions, transportation schedule, weather, and maps. [25] A very useful dataset, commonly combined with AFC data is General Transit Feed specification (GTFS) data. Information in the data includes bus and light rail schedule details, such as stop locations, route details, scheduled arrival and departure times, etc. [43] Some of the technologies commonly integrated in AFC installations are Automatic Vehicle Location (AVL) system, that can dynamically locate the vehicle with just a few meters of error, Global Positioning System (GPS) and Automated Passenger Counter (APC), which is an electronic sensor, used to precisely monitor the amount of passengers getting in and off the transport vehicles. [25] When both AVL and APC are present in a system they are referred to as APC-VL.

2.3.1 Origin Destination Estimation

There are two main AFC ticketing schemes depending on whether the passengers read the smart card in the beginning of their journey (AFC entry-only systems), or both at the beginning and end (AFC entry-exit systems). Bus services worldwide commonly use entry-only AFC systems to avoid delays from the reading of fare media of alighting passengers. These systems involve additional analysis to determine the final destination of passenger journeys, in order to produce the OD matrix [14]. Public transport systems' OD matrices illustrate travel demand, making them vital for transport planning and monitoring. [64] Moreover, in transport networks with multiple operators using a single ticketing system, OD matrices are essential for ticketing revenue management. [31]

Various studies have successfully applied the Trip-Chaining Method (TCM) to infer public transport alighting locations using entry-only AFC data, making it possible to estimate the corresponding OD matrices. [31] A trip chain, also known as a tour, is a series of journeys that begin and end in the same location [36], commonly the traveler's place of residence.

The TCM is a highly complex technique that requires several assumptions, approximations and intermediary steps, and aims to determine the route as well as the precise boarding and alighting locations associated to a trip-leg for each AFC system transaction. [14] These trip-leg structures are primary components of a complete OD trip, therefore, when it is assumed that a passenger uses their smart card twice or more to complete an individual journey, the alighting stop for each trip-leg is estimated, and several trip-legs are chained to form a connected trip. [14] As a result, any imprecision in individual trip-legs alighting stops' estimation are accumulated, and can result in significant errors creating whole OD trips. [3] The TCM cannot be applied to a single transaction because of the lack of linkage to other trips. [46]

This method has been implemented to various transport systems across the globe, as shown by the summary on existing research on Table 2.1. It is based in assumptions, and they vary depending on the transport system and available data. Barry et al. (2002) proposed two assumptions, initially applied to the New York City subway system, and later (2009) to bus, ferry, and tramway systems, in addiction to the subway system, in New York. These assumptions became the foundation for later research in the subject: [13, 14]

1. Most passengers start their subsequent trip from, or near the destination station of their prior journey.
2. Most passengers finish their last trip of the day at the station where they started their first trip of the day.

Following studies have applied similar approaches, introducing additional geographical assumptions. Trepanier et al. (2007) introduced: [79]

3. Passengers do not walk more than a threshold distance between two stops.

For this buffer zone, the same authors used the Euclidean distance of 2000 meters [79]. Zhao et al. considered a shorter Euclidean distance of 400 meters [89] for the Chicago rail network, while Nassir et al. (2011) used a 800 meters boundary in Minneapolis [63]. Wang et al. (2011) and Munizanga and Palma (2012) opted 1000 meters [81, 58, 59], Nunes et al. (2016) used the Euclidean distance of 640 meters [64], and Hora et al. (2017) opted for the Manhattan distance of 3 kilometers [31]. Gordon et al. (2013) applied a 750 meters transfer distance threshold to London smart card and iBus vehicle location data, and suggested that a multi-stage journey should not take an overly circuitous path. To evaluate this the researchers considered a circuity factor of 1,7, calculated by the comparison of the Euclidean distance between the origin and destination stops, and the combination of the Euclidean distances traveled over the correspondent trip-legs. [28] Nassir et al. (2011) also proposed several temporal assumptions:

4. Passengers have a maximum transfer time threshold.
5. Passengers have a minimum time for a certain activity.
6. The estimated alighting time must occur before the time of the following transaction.

Regarding the transfer time threshold, Nassir et al. (2011) considered 90 min, [63] Gordon et al. (2013) considered a maximum time window from 30 to 90 min,[28] while Munizaga and Palma (2012) and Munizaga et al. (2014) defined 30 minutes as the maximum transfer time for a trip. [58, 59] Regarding assumption 5), Nassir et al. (2011) considered a minimum time for a trip activity to be 30 min, [63] while Gordon et al. (2013) defined the minimum time for a transfer between trip-legs to be 5 minutes. [28] Assumption 6) was adopted by Nassir et al. (2011), Gordon et al. (2013) and Alsger et al. (2016) [63, 28, 3] Other common assumptions are:

7. During successive transit trip segments, travelers do not use private forms of transportation. [79, 89, 63, 81]
8. Only stops that have not previously been traveled by the route or direction the passenger boarded are eligible for alighting. [79, 3, 64, 31]

Unlike most studies, Trepanier et al. (2007) implemented an approach to infer single daily trips' destinations, by conducting a similarity pattern analysis for the passenger's records history to estimate the alighting location, provided there was a sufficient amount of data. [79] The authors, also adapted the deduction of the last alighting stop of the day. In the case that the last trip of the day doesn't correlate to the first boarding location of the day, the first the last alighting stop of the day is assumed to be the first boarding location of the following day. [79]

Kumar et al. (2018) applied the TCM to bus, light rail and trains AFC data from the Twin Cities, combined with GTFS, as well as APC-VL. Regarding bus journeys, the system in place is more complex than most, having "regular-routes", and "pay-exit" routes, where the passengers validate their card during alighting, making the passengers' boarding location unknown. The researchers relaxed the common assumptions of the TCM, employing a multinomial logit route choice model to estimate the probability for potential boarding stop location, alighting stop locations, bus delay and passenger's route choice behavior. [43] Using also a probabilistic approach, Assemi et al. (2020) improved the accuracy of the method by proposing a neural network-based method to infer the most likely alighting stop in the TCM. [12]

Yang et al. (2020) proposes a time-dependent hierarchical flow network, implemented using nonlinear programming, that predicts the dynamic OD demand for an urban rail transit system. The model is calibrated using first-order gradients backpropagation, and reassignment of passenger flow using updated weights between different layers, and can integrate more traffic characteristics than traditional methods by addressing factors such as time-dependent features, allowing for accurate predictions regarding travel times spent by passengers departing at each time period. [85] Overall, this research contributes towards improving our understanding about how we can better estimate complex traffic patterns within large-scale subway systems without observed link flows - something which has been challenging researchers worldwide due mainly because most models are limited only to simple intersections rather than realistic congestion levels found throughout these types of transportation infrastructures today.

The main contribution of Lee et al. (2022) is the development of a method for estimating the destination of unlinked trips (trips with unknown destinations), based on previous boarding records and temporal travel patterns from long-term AFC data. The study uses k-means clustering to group passengers together based on travel behavior, and Gaussian mixture models to estimate time-of-day travel patterns within each cluster. [46]

2.3.1.1 Results Validation

Due to the fact that the majority of real-world implementations of AFC systems are entry-only, most of the validation of the previous research has been exogenous, relying mostly on travel surveys, which are not always available. Farzin et al. (2008), Gordon et al. (2013), Munizaga et al. (2014) and Zhao et al. (2007) compared their results against OD matrix surveys available. [26, 28, 59, 89] The dataset used by Zhao et al. (2007) also contained information regarding passenger mile traveled (PMT). [89] Wang et al. (2011), used manual passenger count survey data [81], similarly, Kumar et al. 2018 compared their results against transit on-board survey data. [43] Barry et al. (2002) validated their proposed assumptions (trip-leg transfers occur on the closest stop; last daily destination is near the first), by testing them against travel diary data provided by the New York Metropolitan Transportation Council, concluding that they were correct for 90% of subway users. To validate the results, the study compared the inferred total number of passengers reaching their destination to the actual number of passengers exiting the station throughout the day. Additionally, the researchers estimated the maximum passenger volume at peak times using a trip assignment model. [13] Nassir et al. (2011) conducted spatial validation to determine if the direction of the travels was correctly deduced, using APC-VL data. [63] Similarly, Nunes et al (2017) ruled out implausible alignment locations through the use of AVL data. [64]

Alternately, studies using smart card data from entry-exit AFC systems conduct endogenous validation. These are typically regarded as being more credible. Alsger et al. (2015) and Alsger et al. (2016) used data recorded by the South-East Queensland's AFC system, in Australia, that included information about boarding and alighting times and locations, for buses, trains, and ferries, to validate common assumptions of the TCM. Comparing their results with the real values, they conducted a sensitivity analysis for the transfer time threshold, walking distance threshold, and last daily destination. Regarding the transfer time threshold they concluded that it did not significantly affect passenger travel patterns. However, the results showed that the estimated OD trip matching was slightly better for a transfer time threshold of 30 minutes compared to 60 and 90 minutes. Similarly, regarding the walking distance threshold, an increase over the 800 meters mark did not affect the accuracy of the method considerably. [3] Additionally, for about 80% of the transfer trips, the walking distance was on average 400 m, regardless of the transfer time threshold. [4] Regarding the last daily destination, the researchers evaluated the assumption that it was the same as the first onboarding of the day, against the assumption that it was closest to the first daily boarding station. Algers et al. (2015) concluded that for 82% of passengers, the final stop, was the first daily boarding stop, and 88% of passengers returned to a stop within 800 meters from the stop where they started their first daily trip. [4] Since this estimation causes significant inaccuracies, the authors suggested estimating the final stop to be the closest stop "on the route" to the initial daily boarding location. [3]

Using the same dataset as Alsger et al (2016), Assemi et al. (2020) assessed the distribution of errors at the level of individual stops as well as public transport zones. The results showed that the assumption that passengers disembark at the stop closest to their next boarding point, is not necessarily true for all transactions.

By using a probabilistic approach implemented using neural networks, the accuracy of alighting stop inference can be enhanced. This approach can infer correctly 79.5% of alighting stops that would be incorrectly estimated by the established TCM, and decrease the mean disembark estimation error from 1689 meters to 503 meters. This approach can also increase zoning-level analysis accuracy by more than 5%. [12]

Yang et al. (2020) also conducted endogenous validation, comparing the predicted and actual OD passenger flows, using Root Mean Square Error (RMSE) and Root Mean Square Normalized Error (RMSN). The results of these measures were relatively low for all stations, and the average of the RMSN was below 3%, indicating that the model can accurately estimate passenger flow. [85] Lee et al. (2022) validated their method using data recorded by the Sejong City's AFC system, in South Korea in 2018. The commonly employed TCM successfully predicted the destination of 60.0% of the trips, while the history and pattern analysis based method employed in the study, achieved an accuracy of 74.9%. The method achieved this by effectively linking the destinations of 37.2% of the previously unlinked trips. [46]

2.3.2 Passenger Loads Estimation

Passenger load estimation, refers to the process of estimating the number of passengers, or the level of crowding at a given time on public transportation vehicles, such as buses, trams or trains. It involves predicting the number of passengers who will be on board and/or leave a vehicle or within a specific area, such as a station or platform, based on various factors and data sources.

Table 2.1: Summary of previous research in OD estimation using AFC systems

Reference	Assumptions	Dataset	Validation Dataset	Observation Period	Validation Result
Barry et al. (2002) [13]	1); 2)	6 million subway transactions from New York, USA	Travel diary information	Wednesday in September, 2001	83% of alighting stops inferred (accuracy not determined)
Trepanier et al. (2007) [79]	1); 2); 3); 7) adaption of 2) to take into account next day; history analysis	Bus network transactions from Gatineau, Canada, with GPS information	N/A	July and October, 2003	66% of alighting stops inferred (accuracy not determined)
Zhao et al. (2007) [89]	1); 2); 3); 7)	2.5 million rail and bus transactions from Chicago, USA, with APC-VL information	Household travel survey containing PMT	6-day period in January, 2004	71.2% of alighting stops inferred (accuracy not determined)
Farzin et al. (2008) [26]	1); 2); 3)	505 000 bus transactions from São Paulo, Brazil, with AVL information	Household travel survey (1997)	May 9, 2006	N/A
Barry et al. (2009) [14]	1); 2)	95 million subway, bus, tram and ferry transactions from New York, USA	Travel diary information and household travel survey	2 weeks in April, 2004	N/A
Nassir et al. (2011) [63]	1); 2); 3); 4); 5); 6); 7)	2.17 million bus transactions from Minneapolis–Saint Paul, USA, with APC-VL, GPS and GTFS information	APC-VL data (available for 30% of records)	November, 2008	84.5% to 86.2% alighting stops inferred depending on the parameters (accuracy not determined)
Wang et al. (2011) [81]	1); 2); 3); 7)	Bus transactions from London, England, with AVL information	Manual passenger count survey data	N/A	57.5% to 78.5%% alighting stops inferred depending on the bus route (accuracy not determined)
Gordon et al. (2013) [28]	1); 2); 3); 4); 5); 6)	5 million bus transactions from London, England, with AVL information	Household travel survey	10 weekdays in June, 2011	74.5% of alighting stops inferred (accuracy not determined)

Reference	Assumptions	Dataset	Validation Dataset	Observation Period	Validation Result
Munizaga et al. (2014) [59]	1); 2); 3); 4); 5)	20 million subway and bus transactions from Santiago, Chile, with GPS information	Household travel survey (2010) and OD information from volunteers	1 week in June, 2010	> 80% of alighting stops inferred (accuracy not determined)
Alsger et al. (2016) [3]	1); 2); 3); 4); 6); adaption of 2) to include stops close to 1st boarding	161 446 bus, train and ferry transactions from South-East Queensland, Australia, with GTSF information	Same dataset	March 20, 2013	86.5% for optimal parameters after improvements
Nunes et al. (2016) [64]	1); 2); 3); 8)	3 million bus transactions from Porto, Portugal, with AVL information	AVL spatial validation aspects (no surveys available)	April, 2010	62.4% of alighting stops inferred (accuracy not determined)
Hora et al. (2017) [31]	1); 2); 3); 8)	11 million bus, metro and tram transactions from Porto, Portugal	N/A	January, 2013	N/A
Kumar et al. (2018) [43]	Probabilistic approach	66 374 bus, rail and train transactions from Twin Cities, USA, with APC-VL, GPS and GTFS information	Transit on-board survey (2016)	4 weekdays between March 7 and March 10, 2016	85% of alighting stops inferred (accuracy not determined)
Assemi et al. (2020) [12]	Probabilistic aproach using neural network architecture	159 815 bus, train and ferry transactions from Southeast Queensland, Australia	Same dataset	Wednesday, 20 March, 2013	72.2% alighting stops inferred correctly
Yang et al. (2020) [85]	Time-dependent hierarchical flow network	Subway transactions from Beijing, China	Transactions from Monday, September 10, 2018	September 3 to September 7, 2018	Average RMSN of 2.79, ranging from 1.92 to 3.94
Lee et al. (2022) [46]	Probabilistic approach	2 253 840 bus transactions from Sejong City, South Korea	116 194 transactions from May 28 to May 31	42 weekdays between April 1 and May 31, 2018	74.9% alighting stops inferred correctly

This issue has drawn a lot of attention in transportation studies, as it is crucial public transport systems efficient planning and operation, as well as for accessing passenger comfort measurement and quantification. [73, 87]

Passenger loads estimation address real-time on-board estimated occupancy, contrary to OD estimations studies, that can only be performed on historical data. [87]

Naturally, studies on this type of problem tend to have a big focus on passenger comfort, which is a big concern that gained a lot of importance due to the COVID-19 pandemic. There is an array of different evaluation metrics for passenger comfort, that can be classified as subjective (perception-based, evaluated by surveys) or objective, more easily measured, such as density of standing passengers and proportion of seats occupied on-board. [77]

Overcrowding is particularly an issue during peak periods, since passengers are less willing to sacrifice travel time for greater comfort, assumingly due to commuting time constraints, as found by survey studies conducted by Haywood et. al (2013). [29]

Various techniques and methods can be used for passenger load estimation, including data-driven approaches, statistical models, machine learning algorithms, combinations of the previous techniques, and even simulation-based models, such as the BusMezzo and FastTrips [37] dynamic transit models.

These methods often utilize a combination of different data sources to capture and analyze passenger flow patterns, most commonly they use entry-only AFC, among others.

It is relatively common for ITS to have APC technologies associated. However, due to their high monetary cost, APC systems are often installed solely on a subset of fleet vehicles. [73, 87, 74] Even though APC technologies where not very relevant for OD estimation processes, that focused on passengers, rather than vehicles, for the purposes of crowding estimation, these systems can be very useful. These systems can provide relatively accurate data on the overall passenger flow distribution.

This ground truth information enables the application of supervised learning methods, able to infer the complex relationships between passenger flows and independent variables such as historical demand, dwell time, and headway. Jenelius et al. (2020) applied stepwise regression, lasso regression, and boosted tree ensembles to predict real-time car-specific on-board crowding on the Stockholm metro network using APC and AVL data. The inclusion of real-time data improved prediction accuracy. [34]

Another way these methodologies vary, is in the estimation results. They can be employed for real-time estimations or to predict future demand based on historical data, and some studies estimate values, while others estimate loads, that refer to ranges. Hu et al. (2020) employed a weighted resample bidirectional recurrent neural network, to estimate passenger loads on trains from the mass rapid transit system in the Bay area, connecting several districts. To make the reliable, predictions, a combination of data daily attributes, demand and supply were employed. [32] In the same year, Pasini et. al (2020) used a similar passenger loads model in data from trains in Paris. [66]

Other machine learning algorithms employed for this purpose were the deep belief networks [1], and neural networks [35]. Additionally, Heydenrijk-Ottens et al. (2018) conducted a study comparing random forests, gradient boosting, neural networks, and k-Nearest Neighbors, for the prediction of long- and short-term on-board loading of trams in the Hague, Contrasting machine learning (neural networks, k-Nearest Neighbors) and statistical models (gradient boosting, random forests). The results were promising, indicating the effectiveness of these supervised learning methods. [30]

In the literature, the most commonly employed methodologies for passenger flow estimation are statistical models. In 2018 Hu et al. conducted a study on a bus network in China to generate a crowding profile. They utilized statistical analysis of passenger-flow data, including boarding and alighting events, to estimate the level of crowding on the buses. The study considered factors such as time of day, day of the week, and bus route to analyze the patterns of crowding. Overall, the study demonstrates the effectiveness of statistical analysis in understanding and addressing crowding issues in public transportation systems. [75]

Roncoli et al. (2023) proposed a Kalman filter based passenger load estimation model for the tramway network of Nantes, France. This methodology was conducted and tested using historical and same-day (incomplete) APC and AVL datasets, for passenger comfort level purposes, and was able to estimate the occupancy of a given vehicle, at a given line stop, during a given time interval. The model estimates the number of passengers by calculating the boardings and alighting rates, and applying the assumptions that the passenger flow is homogeneous, in other words that passengers have a equal probability of boarding or alighting at each stop, and that passenger flow is independent from each other. The results suggest that the proposed method is able to deliver good estimation accuracy. [73]

Yin et al. (2022) proposed a set of supervised learning methods to estimate passenger loads in public transport networks, using AFC entry-only data, and APC system's data. This technology was not present on the entirety of the fleet, however it was sufficient to establish ground-truth. Additionally, at some stops, the real information regarding passenger loads was available, recorded by ticket inspectors. This process was developed selecting 3 tram lines in Melbourne, and it was tested on morning peak trips (7:00-10:00) on weekdays between 01/02/2020 to 16/03/2020 in a specific direction. [87]

The paper considers two estimation problems: (1) same-route estimation for routes with APC systems, and (2) cross-route estimation for routes without any APC systems. For problem 1, the models were developed and evaluated using historical observations and independent variables associated with passenger flows, on a particular transit route. Problem 2 involved using a model trained with problem 1, and applying it to a new route with no historical observations. This approach assumes that the relationships between passenger flows and independent variables, established through the trained model, can be generalized to the new route. [87]

The methodology applied a straightforward Extreme Gradient Boosting (XGBoost) to use as a Benchmark. Subsequently, for cross-route estimations, a segment-based estimation method was developed, applying separate methods to different segments of a transit route. This model was developed based on the hypothesis that the relationships between passenger flows and independent explanatory variables are not always consistent across different route segments (based on correlation metrics between boarding or alighting flows). [87]

The results show that the segment-based model led to remarkable improvements in model transferability, due to the features developed from exogenous data sources. This proved that for bus routes, models can be transferred and applied to other transit routes without historical observations. A very interesting conclusion derived from the analysis of the model generated by the supervised learning method, was that for mid-route and downstream stops, boarding passengers do not end up in common destinations, so the boarding flows are not dependent on some common features across different routes. [87]

2.3.3 Limitations

The use of AFC systems' data for modeling purposes has some limitations that have been pointed out by various authors. Firstly, most methods used to estimate information using this data are based on assumptions, which are not always correct, and in most cases cannot be properly validated.

Regarding hardware and software, smart cards usually have a short lifespan, being replaced every few months. Additionally, they can be used by multiple passengers, and a passenger can have multiple cards, since they are easily lost and forgotten. Furthermore, smart card adoption rate is around 80% in most ITS, and fare evasion is still a factor [25] The datasets produced by the AFC inevitably contain errors, some common ones are regarding bus parameters manually produced by the driver. [14, 64, 25] Systems involved in the datasets, such as card readers, AVL systems, GPS, and software can also record incorrect information or even fail. [72]

Chapter 3

Problem Characterization

3.1 Domain

This section aims to establish the context of the problem, by providing an overview of its domain. For this purpose, the public transports' specifications of the Metropolitan Area of Porto (MAP) are presented. In particular, the road transport network, the particularities of the Andante System, which is the AFC system employed in the MAP, as well as the characteristics of the AFC data.

3.1.1 Metropolitan Area of Porto

The second-largest city in Portugal, City of Porto, serves as the focal point of the MAP, a metropolitan area in northern Portugal.

The metropolitan region, which spans 17 municipalities and has a population of about 1,7 million, within an area of 2,040 km2, is the second largest urban area in the nation. [24, 8]

The MAP has an extensive public transportation network that several transport modes, such as rail and road transport, managed by several companies.

The rail transport network includes train transportation, managed by Comboios de Portugal (CP), metro transportation managed by Metro do Porto, and tram transportation managed by Sociedade de Transportes Colectivos do Porto (STCP).

The road transport network includes an extensive bus network managed by 29 private companies, including the biggest player STCP, as well as some touristic buses, taxis, and some ride sharing services. [10]

STCP currently manages 70 bus lines and 2 tram lines, with a route extension of 489km. 59 of these lines operate between 05:30H and 01:00H, and 11 of them operate exclusively during the night, between 00:30H and 06:00H.

The schedules and frequency of the lines vary depending on the day, being different for working days, Saturdays, and Sundays and holidays. Out of all the bus lines, 6 do not operate on Saturday afternoons, Sundays and holidays, and one of them does not operate for the entire weekend and holidays. [78]

The schedules also change during regular season, Summer, and school vacations season.

3.1.2 Andante System

In order to promote the implementation of intermodality in public transport in the MAP, a Complementary Grouping of Companies (Metro do Porto, SA; STCP and CP), with the name of Intermodal Transport of Porto, ACE (TIP), created the Andante ticketing system.

For the organisation and the involved public transportation providers, TIP strives to ensure the deployment of a uniform and exclusive ticketing and pricing system. The consortium is in charge of the revenue distribution model, implementation and management of information, and communication of the infrastructures and modes of transport. [11]

The Andante ticketing system is an entry-only AFC system that employs contactless smart cards with charging points in every station, which is used in a variety of public transport operators in the MAP, such as metro, trains and buses.

This system uses a distance-based fare structure, defined by a zonal structure, made up by geographical travel zones. The cost of a trip is determined by the distance traveled between its origin and destination. During the period of a trip, that is a pre-determined amount of time dependant on the number of zones included in the fare, passengers are allowed to make unlimited transfers, validating the card in the beginning of each journey stage.

There are several fare modalities, such as occasional tickets, temporal subscription tickets, valid for 24 or 72 hours, and most commonly, the monthly subscription.

3.1.3 Anda App

Anda is a smartphone application integrated with the Andante intermodal system, that allows users to travel on public transport in MAP, constituting an alternative to the smart cards. In order to use the app to take a trip, the user must ensure they have NFC, Bluetooth and internet access, open the app, and place the phone close to the validator, similarly to a smart card, and wait for a trip start notification. [69]

The application consists in a check-in/be-out system, meaning that the user only needs to validate the trip while embarking, and the alighting location is automatically detected, and saved, once the costumer leaves the vehicle. The app can be used to follow the trips history, and automatically chooses the best tariff for the trips made. [69]

As this dataset contains complete passengers' trips information, for the purpose of this work, it was used in the validation process.

3.1.4 AFC Data

Every bus and rail stop has a card reader, and every bus is equipped with an AVL system that is fully integrated into the AFC system. The AVL system is used to notify passengers of the next bus stop, and it has been found to carry out that work quite dependably and regularly. It is assumed to assign the stop code to the journey transaction with great accuracy.

A journey includes one or more trip-legs that take place on several routes or in various vehicles and is related to a single fare. In the beginning of each of these journey stages, the user must tap the smart card in a card reader, generating a transaction record, containing attributes regarding the stop where it took place, the travel card used, the time and the location where the journey began, and in the case the reader is inside a bus, it also contains attributes related to the route of the vehicle. Some of these attributes are recorded for on-board ticket inspection purposes, but are useful for the estimation of the destination of journeys.

3.2 PART: 2019 Context

In April 2019 a program called PART, promoted by the environment ministry, was implemented in the Andante system. This program aimed to attract passengers to public transport by supporting transport authorities with an annual grant, enabling tariff reductions. [5]

The implementation of PART brought several changes to the Andante system, starting by an expansion. The AMP became totally covered by Andante, with the addition of operators from municipalities previously non-affiliated. [7] All of the nomenclature from the zonal structure was changed, and some zone limits were re-drawn. Figure 3.1 shows the previous map, from 2017, and Figure 3.1 shows the current map, from 2019. [7]

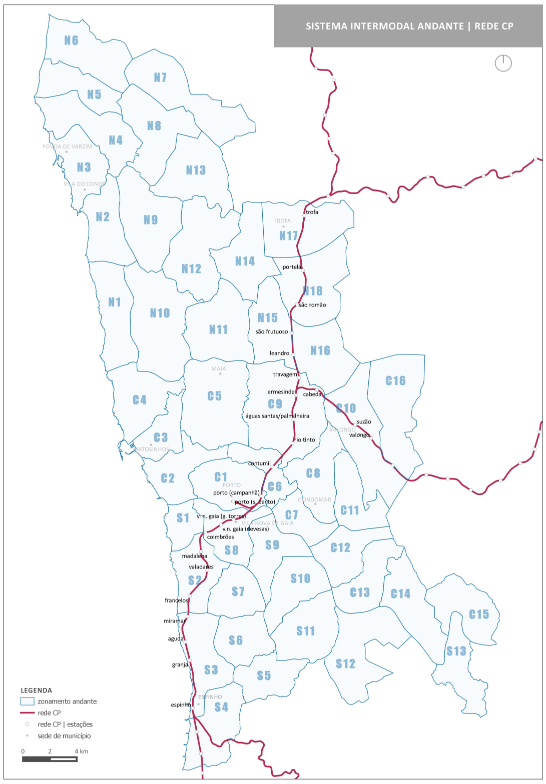


Figure 3.1: Zoning System 2017

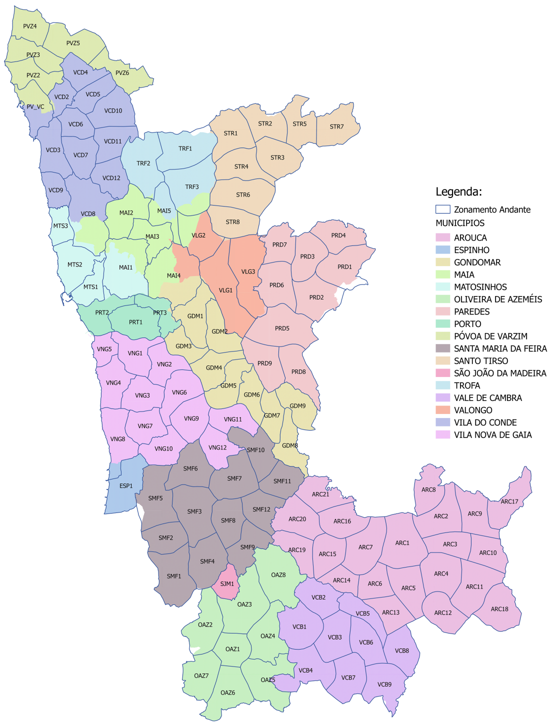


Figure 3.2: Zoning System 2019

The monthly signatures previously varied from Z2, a tariff with access to 2 selected zones, to Z12 a tariff with access to 12 selected zones. The re-structure brought 2 monthly signatures:

Andante 3Z, a tariff with access to 2 selected zones, and Andante Metropolitan, a tariff with access to the entirety of AMP. These signatures came with significant cost reductions for the costumers.

This re-structure of Andante was implemented in April 1st 2019, but was gradual.

3.3 Case Study: Line 205

Once again, the decision was made to study a bus line, since these trips include more information than tram trips, as the validators are located inside the vehicles, rather than in the stations. Line 205 is the most used bus line in the city of Porto. Because of this, both the Andante and Anda datasets contain the most data about this line out of all. Due to this factor the 205 was the line selected for the purposes of this study.

Line 205 has two terminals, Campanhã and Castelo Do Queijo. Campanhã is the biggest Intermodal Terminal in Porto, with connections to trains, trams, city buses as well as national and international buses.

This line's route runs 19 km, 17 km of which constitute the entirety of EN12, also known as Estrada da Circunvalação. This road surrounds the municipality of Porto, bordering Matosinhos, Maia and Gondomar, and plays an important role in the distribution of traffic in the metropolitan area. It is strategically located between the A20-VCI and A4 motorways, functioning as an essential connection between the main road networks and urban areas. This can be observed by its proximity and number of connections with VCI and highway intersections. [9]

This route presents a predominantly urban profile and function. It has a diversity of environments, such as avenues, boulevards and linear squares, which drive a variety of urban and economic activities, such as hospitals, universities, higher education institutions, commercial centers, as well as the intermodal hub of Hospital de S. João. [9]

3.3.1 Stops

Line 205 has a total number of 94 stops described in Table 3.1. Out of these, Campanhã and Castelo do Queijo, the two terminals are common for both directions, as the trips arrive and depart from the same stop in both terminals.

A trip departing from Campanhã to Castelo do Queijo is considered an outward trip (direction 1) and has 49 exclusive stops, with a total of 51 stops with the terminals. The return trip, from Castelo do Queijo to Campanhã (direction 2) has 43 exclusive stops, making a total of 45 stops.

The difference between the stops per direction occurs as a result of locations in one of the directions where it is not logistically possible to stop in the opposite one. Outward trips make a total of 7 stops that do not have direct correspondence to return trips, considering corresponding stops as stops with the same name, and very little distance (the other side of the street typically). Return trips make a single stop not made by the opposite direction.

For the purposes of this study, a stop that does not exist in the opposite direction, was corresponded to the closest stop in the opposite direction.

Table 3.1: Line 205 route

	Direction 1		Direction 2	
Stop Name	**Order**	**Stop ID**	**Order**	**Stop ID**
Campanhã	1	CMP2	45	CMP2
Pinto Bessa	2	PBSS2	45	
Capela S.ra Saúde	3	CSS3	45	
Campanhã	4	CMP4	45	
Noeda	5	NDA1	44	NDA2
EDP	6	EDP1	43	EDP2
Freixo	7	FRX2	42	FRX4
Bonjóia	8	BNJ2	41	BNJ1
Lagarteiro	9	LGT4	40	LGT3
Br.Cerco Porto	10	BCP2	39	BCP1
Pego Negro	11	PNG2	38	PNG1
Bernardim Ribeiro	12	BRB4	37	BRB3
S. Roque	13	SR6	36	SR1
S. Roque (Circunv.)	14	SR5	35	SRGB1
Fábrica do Cobre	15	FBC2	34	FBC1
Vila Cova	16	VLC2	33	VLC1
Parque Nascente	17	PQN2	32	PQN1
Ranha	18	RNH2	31	RNH1
Rebordãos	19	RBD2	30	RBD1
Varziela	20	VRZL2	29	VRZL1
Rot. da Areosa	21	ARSR1	28	
Areosa	22	ARS5	28	ARS4
Enxurreiras	23	ENX2	27	ENX1
Asprela	24	ASP4	26	ASP3
Hosp. S. João (Circunvalação)	25	HSJ12	25	HSJ9
IPO (Circunval.)	26	IPO5	24	IPO4
S. Tomé	27	ST2	23	ST1
Amial	28	AML3	22	AML2
Capuchinhos	29	CPU2	21	CPU1
Via Norte (Circ.)	30	VNC2	20	VNC1
Mte Burgos (Circ.)	31	MTB2	19	MTB1
Congostas	32	CNG2	18	CNG1
Br. Sto. Eugénio	33	BRSE2	17	BRSE1
Quartel Militar	34	QTL2	16	QTL1
Alto do Viso	35	AV2	15	AV1
R. do Senhor	36	RSR2	14	RSR1
Bairro do Viso	37		13	BRVS
S.ra Penha	37	SRP2	12	SRP1
Ruela	38	RUL2	11	RUL1
Rot. A.E.P.	39	RAEP3	10	RAEP2
Preciosa	40	PREC2	9	PREC1
Azenha de Cima	41	AZC3	8	
Lidador /Hospital	42	LDD9	8	LDD6
Quinta Ingleses	43	QTI	8	
Real	44	REAL1	7	REAL2
Teatro Vilarinha	45	TVLR1	6	TVLR2
Parque da Cidade	46	PQC1	5	PQC2
D. Afonso Henriques	47	DAH1	4	DAH2
Dr. Afonso Cordeiro	48	DAFC1	3	DAFC2
Pr. Cid. Salvador	49	PCID4	2	PCID3
Edif. Transparente	50	ETRP2	2	
Castelo do Queijo	51	CQ8	1	CQ8

3.3.2 Variants

Line 205 is the most complex line, presenting multiple variants, that consist in trips that do not travel the entirety of the route. Table 3.2 describes all the the variants, including their number of stops, average duration, as well as the number of planned trips for each, in the year 2019. Figure 3.3 shows the variants in the map.

Figure 3.3: Line 205 Variants

Dir	Code	Departure Terminal	Arrival Terminal	Planed	Stops	Time (min)
1	CMP-CQ	Campanhã	Castelo do Queijo	13 969	51	50
1	SR-CQ	S. Roque	Castelo do Queijo	6 628	39	40
1	SR-LDD	S. Roque	Lidador /Hospital	8 316	30	37
1	MTB-CQ	Mte Burgos (Circ.)	Castelo do Queijo	1 537	21	18
1	MTB-LDD	Mte Burgos (Circ.)	Lidador /Hospital	1 676	12	8
2	CQ-CMP	Castelo do Queijo	Campanhã	14 323	45	45
2	LDD-CMP	Lidador /Hospital	Campanhã	920	38	32
2	CQ-SR	Castelo do Queijo	S. Roque	5 482	36	38
2	LDD-SR	Lidador /Hospital	S. Roque	7 560	29	32
2	CQ-MTB	Castelo do Queijo	Mte Burgos (Circ.)	2 329	19	18
2	HSJ-CMP	Hosp. S. João (Circ.)	Campanhã	314	21	19
2	LDD-MTB	Lidador /Hospital	Mte Burgos (Circ.)	1 512	12	10

Table 3.2: Line 205 Variants

The vehicles running the line 205 alternate irregularly between variants. Figure 3.4 shows all the vehicles routes in a weekday, represented by different colors.

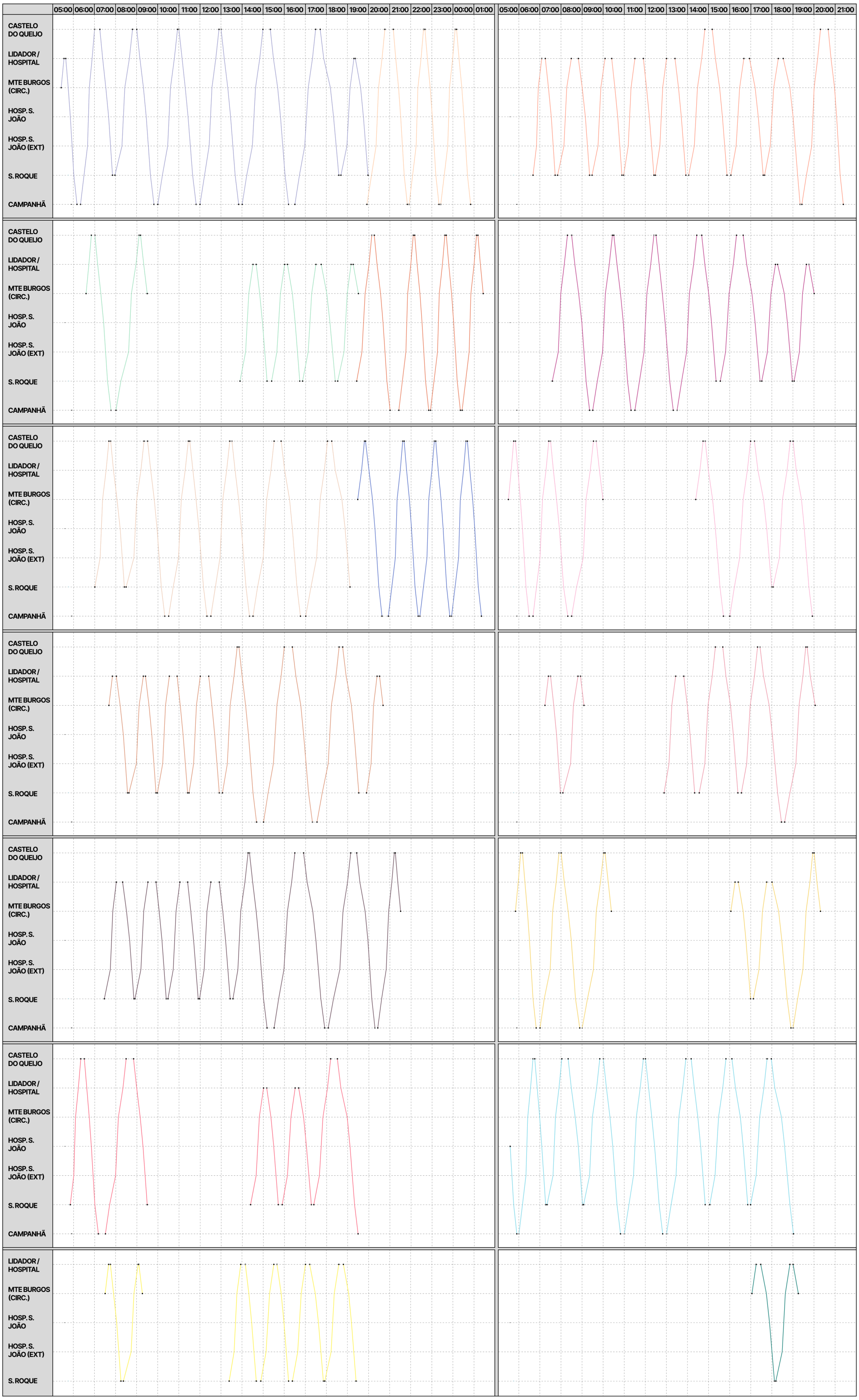


Figure 3.4: Line 205 Vehicle routes in a weekday

Chapter 4

Methodology

This chapter describes the methodology employed to develop the public transport vehicle occupancy estimation model based on data from the AFC system of the MAP. The approach follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a widely accepted framework for data-driven projects. CRISP-DM comprises six key phases: business understanding, data understanding, data preparation, modeling, and finally, evaluation. By adhering to the CRISP-DM framework, a systematic and iterative approach is employed, leading to robust and data-driven insights for the issue in question.

4.1 Technologies

This project was developed in Python, leveraging various libraries and technologies to carry out data analysis, modeling, and visualization. The following are the key imports used throughout the project: pandas, numpy, matplotlib, datetime, sklearn and scipy.

4.2 Data Description

This section provides a comprehensive description of the datasets utilized in the model development process. The data used in this project was relative to the entirety of the 2019 ticketing systems validations.

Understanding the data is fundamental to building accurate and reliable models that can effectively address the problem at hand. An analysis of the datasets' characteristics provide valuable insights that inform modeling decisions.

4.2.1 Andante Dataset

Two partially redundant databases relative to 2019 Andante AFC data were provided. The ROD dataset, with information exclusively related to the road transportation prior to the PART implementation, and information about the entirety of the Andante network after. And the RR dataset, containing information regarding all TIP transactions for all year, including from the app Anda. In both datasets, a line corresponds to the information of a ticket validation.

The ROD dataset had a total size of 27,48 Gigabytes, with a total of 153 858 775 validations, while the RR dataset had a total of 37,44 Gigabytes, and 181 766 001 validations, 698 647 of which were from the Anda app. These were distinguishable because they were the only entries with values for a set of attributes related to the costumer account (*'Customer_Name'*, *'Email'*, *'Phone'*, *'Model'*, *'OS_Version'*, *'Manufacturer'*, *'Brand'* *'Public_Model'*), the end of the trip, (*'Exit_Zone'*, *'ExitDT'*, *'ExitStop'*) and other information relative to bus trips that was not recorded otherwise in the RR dataset (*'Vehicle'*, *'Variant'*, *'Direction'* and *'Line'*). These entries were removed, since the Anda dataset was provided.

Table 4.1 represents all of the attributes from both ROD and RR data. Apart from *'Code'*, the attributes present in both datasets are correspondent. Some attributes have some formatting differences, the *'Validator'* is saved as an integer in ROD and as a hexadecimal in RR, and the *'ValidationDT'* in ROD includes time information up to the second, while in RR it only includes up to the minute.

The ROD dataset is more complete, including specific information regarding road transportation, namely the *'Line'*, *'Direction'*, *'Variant'* and *'Vehicle'* of the validations. It also includes the *'TripBeginningDT'* column, that stores the information of the date and time relative to the beginning of the bus trip in the departure terminal. All this information is useful for the purpose of this work.

However, the RR dataset exclusively has the *'ExitZone_RR'*, that consists in the exit zone, estimated by TIP, for the purposes of fare division between operators. This column could be useful for the estimation of passenger flow, making both datasets essential.

Some of the columns are categorical, including *'ProviderCode'*, *'DiscountCode'*, *'TarifCode'*, *'Zoning'*, *'Direction'* and *'Variant'*. The *'DiscountCode'* provides information regarding the passengers' social profile, e.g. 'Student', 'Senior', etc. The *'Zoning'* has two possible values: '2' to indicate the old zoning system and '5' for the new zoning system. Similarly, the *'Direction'* has the value of '1' for outward trips, and '2' for return trips, while the *'Variant'* is '1' for regular trips, '2' for night trips and '3' for circular trips (with a circular route).

4.2.2 Anda Dataset

The Anda dataset, with information relative to the trips performed using the Anda app in 2019, was provided by TIP. This dataset consists in a table with a total of 700 446 rows, representing attempts to initiate trips using the app, and 33 attributes, containing several information about the trips, including boarding and alighting. Because of this, it was used in the validation process of the developed model. The attributes related to the customer account are not relevant for the purposes of the work, namely *'UserHash'*, *'Manufacturer'*, *'Brand'*, *'Model'*, *'OS_Version'* and *'Public_Model'*.

Table 4.2 describes the remaining Anda attributes. As explained in the previous chapter, passengers are allowed to make unlimited transfers during the pre-determined period of a trip. The

Attribute Name	**Description**	**ROD Data Type**	**RR Data Type**
Code	Validation ID	int	int
CARD	Card ID	hex	hex
Validator	Validator ID	int	hex
ProviderCode	Code for the service provider	int	int
DiscountCode	Code for the discount type	int	int
TarifCode	Code for the type of ticket	int	int
Zoning	Code for the zoning system	int	int
EntryZone	Zone ID for the validation	str	str
ValidationDT	Timestamp for the validation	datetime	datetime
ExitZone_RR	Zone ID for the estimated exit	–	str
EntryStop	Stop ID for the validation	int	str
Line	Line ID	int	–
Direction	Direction code	int	–
Variant	Variant code	int	–
Vehicle	Vehicle ID	int	–
TripBeginningDT	Timestamp for the beginning of the trip	datetime	–

Table 4.1: ROD and RR Attributes description and characteristics

Anda app saves the information from all trip legs, including a common identifier *'Trip'* for the trip they belong to, and the date time, zone, as well as stop name for the beginning and end of the trip (*'Trip Beginning DT'*, *'Trip End DT'* , *'Trip Entry Stop'* , *'Trip Exit Stop'* , *'Trip Entry Zone'* and *Trip Exit Zone*). While this information is not relevant for the purpose of occupancy estimation the *'EntryZone'* column was kept because it is referenced in *'Trip Leg Entry Zone'*, that can have *'BeggininigOfTrip'* as a value.

The *'Client Group'*, *'Unique Stop'*, *'No Payment Method'* and *'Processing Result'* attributes are auxiliary to the identification of trip attempts that did not go through.

An issue with the Anda dataset worth mentioning is that it only contains stops names, that for some cases do not work as unique identifiers.

4.2.3 Auxiliary Datasets

In addiction to Andante and Anda, a number of auxiliary databases were used to better understand and enrich the data.

Table 4.2: Anda Attributes description

Attribute Name	**Description**
Client Group	Indicates whether the trip was made by a costumer or if it was a system revision test
Trip	Trip ID
Trip Beginning DT	Timestamp for the beginning of the trip
Trip End DT	Timestamp for the end of the trip
Trip Entry Stop	Name of the trip entry stop
Trip Exit Stop	Name of the trip exit stop
Trip Entry Zone	Zone ID for the trip entry
Trip Exit Zone	Zone ID for the trip exit
Unique Stop	Indicates whether a the trip had more than a single stop
No Payment Method	Indicates whether there was a valid payment method
Finishing Mode	Descriptor for the mechanism that triggered the identification of the end of the trip
Trip Leg	Trip leg ID
Operator	Name of the trip operator
Vehicle	Vehicle ID
Line	Line public name
Direction	Direction describer
Trip Leg Beginning DT	Timestamp for the beginning of the trip leg
Trip Leg End DT	Timestamp for the end of the trip leg
Trip Leg Entry Stop	Name of the trip leg entry stop
Trip Leg Exit Stop	Name of the trip leg exit stop
Trip Leg Entry Zone	Zone ID for the trip leg entry
Trip Leg Exit Zone	Zone ID for the trip leg exit
TariffGroup	Descriptor for the type of ticket
SocialGroup	Descriptor for the type of discount
Zone nr	Number of zones that have been travelled using the card
Zone list	List of the zones associated to monthly tickets
Processing Result	Indicates weather the trips was processed and charged or not

Auxiliary tables were provided with complementary information regarding the codes of the categorical attributes *'ProviderCode'*, *'DiscountCode'* and *'TarifCode'*. Furthermore, a listing of the lines with information regarding the correspondence between the internal system IDs and the public nomenclatures was used, along with a listing of the correspondence between the old and the new zoning system, for the zones that were not re-drawn. Additional information about the lines was extracted from the STCP website.

The last dataset used, was featured the full details of the Andante stops, including the name, the public nominal identifier, the internal numeric identifier, as well as geospatial information, in two different formats. The first, SAE_X and SAE_y, two floating-point numbers representing the stop's X and Y coordinates in the Sistema de Apoio à Exploração System (Exploration Support System), which is a system for managing and monitoring public transportation in Portugal. The second, UTM_X and UTM_Y, representing Universal Transverse Mercator coordinates. Using the pyproj python library, the UTM coordinates were converted to latitude and longitude coordinates.

4.3 Data Preparation

The data preparation phase is a critical step in any model development project, as it ensures that the data used to build the model is accurate, complete, and properly formatted. The following chapter describes the steps performed in the data preparation phase.

4.3.1 Data Selection

For the purposes of this study, only STCP data was used, which corresponds to the *'ProviderCode'* 3. ROD included 69 780 478 STCP validations, RR had 71 558 942, and Anda had 232 957. After this selection, the ROD *'EntryStop'* attribute was entirely NULL, as the information about the validation's entry stop was only present for validations made in stations, and not in buses.

The considerable difference between the number of records in ROD and RR is due to duplicate entries in the RR dataset. These duplicates have the same *'Code'*, and are generated when the system makes more than one prediction for the *'Exit_ZoneRR'* for a trip. This occurrence was particularly high in the month of April, following the implementation of PART.

4.3.2 Data Integration

The ROD and RR datasets were combined into a single Andante dataset. To find the matching validations between them, tuples of the *'CARD'*, *'ValidationDT'*, and *'Validator'* attributes were used. In order to do this, some conversions were necessary, as the columns *'ValidationDT'* and *'Validator'* were in different formats in both datasets (see Table 4.1). The *'Validator'* from ROD was converted from int to hexadecimal in order to match the RR format. Regarding the *'ValidationDT'*, an auxiliary column was created in ROD, with a copy of the datetime of the validation with the information truncated at the minute (second set to zero). Using this method it was possible to match the datasets without losing accuracy in the information.

Some groups of 'CARD', *'ValidationDT'*, and *'Validator'*, representing validations made by the same card, in the same validator, in a period of one minute, had more than one validation. Because of this, they could not be used as unique identifiers for a one-to-one match. As a solution they were considered duplicates, and the last validation in these cases was kept, since the first one had a higher chance of having errors.

Information exclusive to the ROD dataset included *'Line'*, *'Direction'*, *'Variant'*, *'Vehicle'*, *'TripBegginingDT'* and *'EntryStop'*, attributes considered crucial for the purpose of the study.

The only information present in RR but not in ROD was regarding the *'ExitZone_RR'*. Since this information wasn't deemed essential for the model, the final Andante dataset matched the ROD and the RR datasets, and included the validations from ROD without a RR match. This dataset had a total of 69 633 212 validations, 295 892 of which had no information regarding the *'ExitZone_RR'*.

Table 4.3 shows the impact of the Data Integration process by month. The removal of validations made by the same *'CARD'* in a specific *'Validator'* during the same minute period resulted in a final dataset with 2,69% less validations that the biggest original one (RR).

Table 4.3: Data Integration's impact on the datasets

Month	**RR anda Removed**	**Duplicates**	%	**ROD**	**Duplicates**	%
1	5 286 729	17 860	0,34	5 297 845	14 196	0,27
2	5 156 418	14 365	0,28	5 170 809	13 633	0,26
3	5 584 420	19 760	0,35	5 597 031	15 101	0,27
4	7 456 169	1 948 994	26,14	5 541 477	11 844	0,21
5	6 585 835	47 733	0,72	6 590 740	13 894	0,21
6	5 585 019	23 760	0,43	5 598 728	11 895	0,21
7	6 235 822	25 612	0,41	6 250 755	13 835	0,22
8	5 031 406	15 314	0,30	5 050 582	12 010	0,24
9	6 110 914	29 838	0,49	6 119 099	11 886	0,19
10	6 838 715	34 658	0,51	6 855 269	11 269	0,16
11	6 107 788	20 101	0,33	6 120 323	8 650	0,14
12	5 579 707	23 627	0,42	5 587 820	9 053	0,16
Total	**71 558 942**	**2 221 622**	**3,10**	**69 780 478**	**147 266**	**0,21**

4.3.3 Data Cleaning

Data cleaning consists in finding and fixing any flaws or inconsistencies in the data, such as missing values, outliers, or duplicate records.

A total of 1510 validations had values for the *'Line'* attribute that were not present in the STCP listing. These were considered errors and removed.

The dataset also included some validations labeled as regular, or circular, *'Variant'* wise, dated on the '2019-01-01' between 00:00H and 01:00H, referring to trips that started in '2018-12-31'. These 880 validations were removed.

4.3.3.1 Removing Duplicates

Duplicate records are commonly generated by passengers checking how much trip time they have remaining, as well as passengers that don't remember tapping the smart card.

As a result of the Data integration process, duplicate validations occurring within the same minute were already removed from the dataset. Since the data refers exclusively to bus trips validations, and for the Andante system each passenger needs their own card, duplicates were defined as validations made by the same *'CARD'*, in the same bus trip.

The *'TripBginningDT'* was used to identify validations belonging to the same trip, in conjunction with the *'Line'*, *'Direction'* and *'Vehicle'*, as well as the *'Date'* and 'Hour' components of the *'ValidationDT'*. This strategy was chosen since the *'TripBeginningDT'* attribute alone was not sufficient to use as a unique identifier. Moreover, this attribute is generated manually by the bus drivers in the process of changing the direction of a route, upon completion of a trip. When this process was forgotten or not done correctly, multiple timestamps from the date '2000-01-01' were generated, making it likely to mistakenly identify validations with illogical values, as being part of the same trip.

4.3.3.2 Removing Outliers

The various fares offered by the Andante system include a number of monthly signatures, one-time tickets, and tickets with 24- or 72-hour validity, marketed toward tourism. In order to identify outlier passengers, an analysis was conducted regarding the amount of trips made by passengers within the rate types previously mentioned, as different types of tickets presented different travel patterns.

In order to do this, the *'TarifCode'* attribute was used. The passengers with *'TarifCode'* 2 (Subscription), 140 (Andante 3Z), 141 (Municipal Subscription), 142 (Andante Metropolitan) and 152 (AMP/CP Subscription) were grouped together as monthly subscription passengers, passengers with *'TarifCode'* 6 (Andante 24h), 19 (Andante Tour-1) and 20 (Andante Tour-3) were grouped together as temporary subscription passengers, and the *'TarifCode'* 1 (Single Ticket) was used to identify occasional passengers.

For each category, the validations were grouped by passenger, using the *'CARD'* attribute, and statistics were calculated on the amount of bus trips per day, and the number of days that passengers travelled in 2019.

A guideline for the outlier criteria was determined by adding 5 times the daily validations standard deviation, to the mean value of daily validations, since it is assumed that values significantly higher than this guideline are unusual and may indicate abnormal behavior or errors in the data.

	Monthly Subscriptions	**Temporary Subscriptions**	**Occasional Tickets**
Validations	59 292 535	533 070	9 614 300
Unique CARDs	358 544	145 457	1 058 515
Avg val per CARD	165.37	3.66	9.08
Avg daily val	2.35	2.20	1.59
Std daily val	1.51	1.40	0.83
Avr + 5 * Std	9.89	9.21	5.73
Outlier criteria	10	10	7
Avr days per CARD	70.47	1.67	5.70
Nr outlier CARDs	145	14	19

Table 4.4: Outliers identification criteria and results

Based on this calculation, if the count of validations for a passenger was equal or greater than outlier criteria, on more than the average number of days travelled by passengers of their group, that passenger was considered an outlier. Table 4.4 shows the criteria and results of this process.

By taking into account both the count of validations and the distribution across multiple days, this approach identifies passengers with consistent patterns of abnormal behavior as outliers. All of the validations made by this passengers were removed from the dataset.

4.3.3.3 Missing and Illogical Values

One crucial step in working with a dataset is to identify and handle missing values appropriately. Missing values can occur due to various reasons such as data collection errors, system failures, or intentional omissions. It is essential to understand the extent and patterns of missing data to ensure the integrity and reliability of the subsequent modeling process.

Upon inspecting the dataset, the following variables exhibit missing values:

'ExitZone_RR' had the most missing values, 186 558 corresponding to 0,27% of the dataset. These were validations where the system was not able to estimate the exit zone, or validations from the ROD dataset with no RR corespondent. As previously determined, the absence of this attribute do not present a problem.

The most frequent cause of missing and illogical attribute values of the Andante data is the direction changeover between consecutive vehicle trips. This process must be manually signified by the driver. Passengers can board the bus before the driver signals the completion of the previous trip, creating a record of a validation in the line terminal, which is illogical. If passenger board the bus after the driver signals the completion of the previous trip, yet before the next trip is initiated, it could originate missing values for the *'TripBegginingDT'* and the *'Direction'*. *'TripBegginingDT'* had 50 660 missing values corresponding to 0,073% of the dataset, while *'Direction'* had 616 missing values corresponding to only 0,001% of the dataset.

Missing *'EntyStop'* are likely the result of a communication failure of the AVL system, 0,012% of the dataset (8580 entries) was missing this attribute.

After examination of the frequency of values for the dataset's attributes, and some cross-examination, it was concluded that the Andante system uses 0 as a NULL value for numeric attributes. A total of 314 871 validations had *'Vehicle'* set to 0, making 0,456% of the data, 15 018 had *'EntryStop'* set to 0, totalling 0,022% and 90 513 entries, 0,131% of the data had *'Validator'* set to 0.

The column with the most illogical values was *'TripBegginingDT'*, with 1 991 023 timestamps with dates prior to 2019, from mostly 2000, making up 2,88% of the dataset.

The process of mitigating missing and illogical values was performed exclusively to the data from the line 205, as it is the data from the case study. This process was complex and used specific information to the line, extracted manually from the STCP website. Additionally, it has a high execution time, and since the rest of the data was not used for the development of the model, it was not deemed necessary to mitigate the missing and illogical values. The mitigation process is described in section Line Concept, and can be applied to other lines.

4.3.3.4 Results

The data cleaning process started with the integrated Andante dataset, and resulted in a 0,82% reduction of the validations, from 69 633 212 to 69 063 170. Table 4.5 shows the impact each phase had in this process, divided by month. The total reduction of the dataset, accounting for the duplicates removed in the Data Integration was of 3.8%.

Table 4.5: Data Cleaning results by phase

Month	Andante	Duplicates	%	Outliers	%
1	5 283 649	15 166	0,29	24 125	0,46
2	5 157 176	14 447	0,28	24 328	0,47
3	5 581 930	17 547	0,31	27 131	0,49
4	5 529 633	15 357	0,28	34 246	0,62
5	6 576 846	18 666	0,28	37 859	0,58
6	5 586 833	16 036	0,29	33 747	0,60
7	6 236 920	18 515	0,30	37 136	0,60
8	5 038 572	16 057	0,32	32 906	0,65
9	6 107 213	16 707	0,27	32 621	0,53
10	6 844 000	17 399	0,25	33 526	0,49
11	6 111 673	13 952	0,23	30 218	0,49
12	5 578 767	13 458	0,24	28 892	0,52
Total	**69 633 212**	**193 307**	**0,28**	**376 735**	**0,54**

4.3.4 Data Transformation

The data transformation process begins with the initial dataset, described in Table 4.1. The first step is to perform feature selection to identify the most relevant features for the purpose of estimating the occupancy of trips, and eliminating redundant or irrelevant features, reducing dimensionality. The *'Code'* was removed, as pandas' DataFrames have by default a unique integer index to each row. Since the *'ProviderCode'* 3, corresponding to STCP was previously selected, this feature is no longer useful. The *'Zoning'* column was also removed as it is not relevant for the work, and since the *'Variant'* can be deduced from the *'Line'* feature, it was also eliminated.

Using the provided listing of correspondence between the old and the new zoning systems, the *'EntryZone'* and *'ExitZone_RR'* columns were updated to the new zoning system.

Additional relevant features were generation using feature engineering. Firstly, a *'Date'* column was generated using the *'ValidationDT'* and the *'Line'*. For nocturnal lines, operating between 00:30H and 06:00H, the date was directly extracted from the Validation datetime column, for the rest of the lines, operating between 05:30H and 01:00, if validations occurred before 03:00H, the date was set to the previous day, otherwise it would be set to the validation day.

Using the accurate *'Date'* from the validations, as well as a list including the dates of the holidays in Portugal in the year 2019, the *'WeekDay'* was generated. This column identifies weather the entries were from a work day, a Saturday, or a Sunday/holiday, since these are the types of days with different schedules and trips frequency.

A *'Month'* and an *'Hour'* attributes were also added, using the *'ValidationDT'* for seasonal and trends analysis.

The transaction records were ordered chronologically.

4.4 Line Treatment

Line 205 was the line in 2019 with the most passengers, with a total of 3 351 681 validations.

The only information regarding the line present in the datasets provided was that its internal ID is 16. The information regarding the line route and variants was extracted from the STCP website. This consisted in two tables, one for each direction, with the route ordered. These tables included the name (*'StopName'*), public nominal identifier (*'StopID'*), and zone of each stop (*'ZoneID'*). A *'Order'* identifier, starting at 1 and incrementing by 1, and a *'Direction'* columns were added to both tables, that were subsequently joined.

Using the provided Andante stops dataset, the table was completed with the internal numeric identifiers of the stops (*'StopNr'*), correspondent to the *'EntryStop'* column, and their geographical coordinates (*'StopLat'* and *'StopLong'*).

Using this new Line 205 table, the Andante validations were completed with the information regarding the stop *'Order'*. This attribute is very helpful for better readability of the data, and while it is not itself a unique identifier, used in combination with *'Direction'* it can act as one.

4.4.1 Mitigating Illogical Values

During the development of a model using a dataset, it is crucial to address and mitigate any illogical values present. Illogical values can significantly impact the accuracy and reliability of the model's predictions. In this section, we outline the steps taken to identify and mitigate illogical values for the various attributes in the dataset.

4.4.1.1 Vehicle

Vehicle IDs from the STCP bus fleet are visible in the buses and consist in numbers of 4 digits. The Line 205 had 10 644 validations that did not fit this criterion.

In order to complete these entries with the correct vehicle ID information, the *'Validator'* was used. If there were other validations from the same day made in the same *'Validator'* with a valid *'Vehicle'*, that identifier was added to the incorrect entry. Following this process only 798 validations with incorrect *'Vehicle'* remained. These validations were distributed across 4 non-related days. After analyzing their distribution with respect to *'Direction'*, and *'TripBegginingDT'*, it appeared that the remaining validations with invalid *'Vehicle'* correspond to a single bus for each *'Date'*, therefore these entries were not eliminated.

4.4.1.2 Validator

There was a total of 7675 validations with *'Validator'* ID 0. These were analysed and did not appear to have any errors. Since the *'Validator'* information is not relevant for the purpose of this work, they were not removed.

4.4.1.3 ExitZone

The *'ExitZone_RR'* column is the only column with missing values, with a total of 10 311 missing entries, making up 0,301% of the dataset. This column was analysed for illogical values, considering values of zones that were not part of the route of the bus between the *'EntryStop'* of the validation and the terminal. A total of 1 619 793 of entries fit this description, making up a total of 48,37% of the validations. Taking this information into consideration, it is very likely that this column represents the estimated exit zone of the entirety of the trip, rather than the trip-leg. Since this column is the only information regarding the exit of the passengers present in the dataset, there is no way to mitigate the incorrect *'ExitZone_RR'*, so the values of this columns were assigned to NULL. Moreover, the quantity of illogical values for this column puts into question the accuracy of the estimation of all exit zones, therefore it was decided that this information would not be used for the development of the model.

4.4.1.4 Stop

There are many possible illogical values regarding stops, some examples are validations in the line's terminals, AVL failures, or validations from stops not present in the bus route.

In the dataset 2951 entries had *'EntryStop'* numbers not present in line 205. 2908 of them had the same non-existing stop, and had information regarding Castelo do Queijo-Campanhã trips that took place between 02:58H and 06:20H, in the week from 5 and 12 of May. These trips were direct (with no stops) and took place as a sporadic transport reinforcement, in support of an academic event, they were therefore removed. [68]

After finding these non-planed trips, the rest of the dataset was analysed in search of validations from trips taken between 02:00H and 04:00H, during not programmed schedules. An additional 226 validations were removed from the dataset.

From the remaining validations, 40 had stop code 0, all from the same bus, day, and with the same *'TripBegginingDT'*, corresponding to an entire trip with AVL failure, and the other 3 entries corresponded to Andante stops outside the 205 route, all with invalid *'TripBegginingDT'*. These entries made up all the validations of their corresponding *'TripBegginingDT'*, therefore it was not possible to complete the incorrect information using the remaining validations from the same trips, so they were removed.

As for terminal validations, 20 143 entries with *'Direction'* 1, and *'EntryStop'* code from Castelo do Queijo were identified. For these validations, the *'ExitZone_RR'* was deleted, as it was probably incorrectly calculated, and the *'Direction'* was changed to the opposite one. There is no need to update the *'TripBegginingDT'* column, as the grouping of validations by trip is not going to be based on it. The same process was applied to the opposite *'Direction'* on a total of 27 066 validations. Validations in terminals for the other variants were not possible to identify at this point.

4.4.1.5 TripBegginingDT

For the year of 2019 there was a total of 64 566 trips programmed for the line 205, described in Table 3.2. The Line 205 dataset however, contains 101 593 different *'TripBegginingDT'*, corresponding to a total of 157,35% of the scheduled trips. Out of all the *'TripBegginingDT'* values, 18,65% of had a single validation, 32,97% had between 1 and 3 validations, and 49% had validations in a single stop. Additionally, 17.57% of the existing *'TripBegginingDT'* consist in several timestamps from the date 2000-01-01.

Is is worth noting that even though a *'TripBegginingDT'* doesn't always correspond to a trip, validations with the same *'TripBegginingDT'* value seem to be consistently part of the same trip.

This attribute depended on a manual process, explaining why it was not as reliable. After analysis of entire days for certain vehicles with exclusively invalid *'TripBegginingDT'*, several errors generated by the the lack of trip initialization were identified. It was common that the new trips would be signified by the system (with a new *'TripBegginingDT'*) every few stops. In some cases the clock of the two bus validators was not tuned, resulting in consecutive validations shifting between prior stops. Some validations in the middle of the route would incorrectly have the departure terminal *'EntryStop'*, or another random stop associated. These cases with clearly incorrect *'EntryStop'*, most likely due yo AVL system malfunctioning, were always signified by a new *'TripBegginingDT'*, and were often the only validations in those "trips".

In order to identify and to correct errors of this sort, the the validations were grouped by *'TripBeggingingDT'*, and the resulting dataset was ordered by *'Date'*, *'Vehicle'* and *'ValidationDT'*, in a manner where all validations groups from a certain vehicle appeared consecutively. Using this structure, consecutive trips done in the same day by the same bus in the same direction were identified. If the last stop from the first trip was further into the direction that the first stop of the second trip, and the average time between stops, during this segment was inferior to 35 seconds, it was assumed that the first validation from the second group was an error. This was checked by evaluating both groups, and *'EntryStop'* from the error validations would be set to the *'EntryStop'* of the validation closest in time.

4.4.2 Grouping Data by Trip

To group the data by trip, a new *'TripID'* was created. To do this, the dataset was ordered in a manner where all validations from the same trips appear consecutively. Considering the problem at hand, this was done by ordering the validations firstly by *'Date'*, for the same dates by *'Vehicle'*, and for the same vehicles by *'ValidationDT'*.

The *'TripID'* was an incremental int starting at 1, and was created by going through the dataset and comparing consecutive pairs of validations. If any of the following conditions were met, it was considered that a change in trips had occurred:

1. There was a change in *'Date'*.
2. There was a change in *'Vehicle'*.
3. There was a change in *'Direction'*.
4. There was a change in *'TripBeggingingDT'*, and the time period between the *'ValidationDT'* of the current row and the previous row was more than 50 minutes.

The last condition was used to identify two consecutive trips in the same direction as two trips instead of one. Considering the bus routes in Fig. 3.4, this should not happen, however the dataset is incomplete. The *'TripBeggingingDT'* was used for this purpose because even though a *'TripBeggingingDT'* doesn't always correspond to a trip, validations with the same value for this attribute seem to be consistently part of the same trip. In other words, every new trip is signaled by a new value for *'TripBeggingingDT'*.

One more condition was evaluated, adapted depending if the validation had a valid *'TripBeggingingDT'* (from 2019) or an invalid one (from 2000-01-01) indicating a higher probability of errors and inconsistencies in the validations. For validations with valid *'TripBeggingingDT'* a new trip was identified if there was an *'Order'* decrease from one validation to the next.

For validations with invalid *'TripBeggingingDT'* it was only considered that a new trip had started if there was an *'Order'* decrease superior to 6. This tolerance was used to take into account the errors that these trips often present.

This process resulted in the identification of 53 706 different trips, corresponding to 83,18% of the planned ones.

4.4.3 Trip Concept Definition

The goal of this work was to estimate the occupation of the vehicles using the data from the ticket validations. In order to do this, the data had to be re-structured, condensing the information from a trip in a single row.

With the *'TripID'*, the data from the different trips could be easily grouped. For each *'EntryStop'* a list with the number of passengers, first and last *'ValidationDT'* was created. This information was joined in a python dictionary. These structures are an unordered collections of comma-separated key-value pairs, used to store and retrieve data efficiently. Each key within a dict must be unique, and it is used to retrieve the corresponding value associated with it. This dict, called *'StopsInfo'* had a key-value pair for each stop made by the trip, with the key being the *'Order'* from the stop and the value being the list described earlier.

A *'StopsInfo'* instance looks like { *'Order'* : [*'nr_passengers'*, *'First ValidationDT'*, *'Last ValidationDT'*], ... }

A new database called unique_trips was created with all of the trips from line 205 taken in 2019 with the following information: *'Date'*, *'Vehicle'*, *'TripID'*, *'Direction'*, *'StopsInfo'*, *'TripBegginingDT'*, *'First ValidationDT'*, *'First ValidationDT'*, *'FirstOrder'*, *'LastOrder'*, *'nr_stops'*, *'nr_passengers'*, *'Hours'*, *'WeekDay'*, *'Duration'*.

4.4.3.1 Variant Identification

The variant of the trips is an important information for the development of the model.

To help identify the variant, the departure and arrival terminals were estimated using the stops of the first and last validation of the trip accordingly (*'FirstOrder'* and *'LastOrder'*). Depending on the *'Direction'*, the departure and arrival stops were estimated assuming that no trips traveled between terminals with no validations. For example in the *'Direction'* 1, that travels the terminals in the order: Campanhã (CMP); S. Roque (SR); Lidador/Hospital (LDD), Mte. Burgos (Circ.) (MTB); Castelo do Queijo (CQ); if the first validation of a trip was between SR and LDD, it was assumed that the trip departed from SR, even though it was possible that it departed from CMP. Then, trips departing from CMP were assigned the arrival terminal CQ, as that is the only variant departing from CMP (see Table 3.2).

For consecutive trips in opposite directions, the arrival terminal of the first, is always the departure terminal for the second, as seen in Fig 3.4. Taking this into account, for consecutive trips in the same *'Day'*, done by the same *'Vehicle'*, with opposite *'Direction'*, and a change period inferior to 01:30H (minimum time assumed to perform a round trip), the arrival terminal of the first, and the departure terminal of the second were compared. If they did not match, they were adjusted accordingly. Let's take two trips that fit the previous criteria as an example, if the first was assumed to travel SR-CQ, and the following LDD-CMP, the second trip's departure stop would be adjusted to CQ instead of LDD, as departing from LDD would imply that the bus travelled between CQ and LDD with no passengers.

This results from this process can be consulted in Table 4.6, the first five rows correspond to the Direction 1 variants, while the rest of the table refers to direction 2.

Code	Planed	Identified	%
CMP-CQ	13 969	13 714	98,17
SR-CQ	6 628	7 262	109,57
SR-LDD	8 316	5 626	67,65
MTB-CQ	1 537	170	11,06
MTB-LDD	1 676	7	0,42
CQ-CMP	14 323	12 446	86,90
CQ-SR	5 482	7 831	142,85
CQ-MTB	2 329	138	5,93
LDD-CMP	920	770	83,70
LDD-SR	7 560	5 492	72,65
LDD-MTB	1 512	19	1,26
HSJ-CMP	314	231	73,57
TOTAL	64 566	53 706	83,18

Table 4.6: Variants Identification in Andante

Some variants had a number of identified trips higher than the actual number of planned trips. This error is most likely due to validations with wrong *'EntryStop'* information, that mislead the route of the bus leading to incorrect variant identifications.

4.5 Anda Treatment

The initial Anda dataset contained 700446 client trip legs, entries with the *'Operator'* STCP were selected, making a total of 232957. Entries from customers corresponded to the value 'Others' in the *'Client Group'* attribute, the *'Processing Result'* was used to identify 'Trip not broken', and the *'Unique Stop'* to select trips that has more than a stop (with value 'No'). The entries with a value of 'Beginning of Trip' for the *'Trip Leg Entry Stop'* were completed using the *'Trip Entry Stop'* and rows with *'Trip Beginning DT'* equal to *'Trip End DT'* were deleted. After this process the dataset was left with 186427 trip leg entries.

These were completed with *'Date'*, *'WeekDay'*, *'Hour'* and *'Month'*, using the same strategies employed for the Andante dataset. The information of zone attributes was also updated to the new zoning system in the same way. The *'Duration'* of passengers trips was calculated with the entry and exit timestamps.

At this point, the 10413 trips from line 205 were selected, and their *'EntryOrder'* and *'ExitOrder'* were added. The data had 20 instances were the order of the entry station was higher than the order of the exit station. Since the system is check-in/be out, it was assumed that the *'Trip Leg Exit Stop'* were incorrectly calculated. It is worth to note that 18 out of the 20 illogical stops were terminals, 9 Castelo do Queijo, 6 Campanhã, 2 S. Roque and 1 Lidador/ Hospital.

Using the table of line 205 trips, the information regarding the *'TripID'* and the *'Variant'* was added to the dataset.

4.6 Data Analysis

In this section, some helpful insights from the data analysis process will be presented. This process was conducted with the objective of better comprehending the data and its patterns and trends, in order to inform the model development process. The data analysed was exclusively from line 205.

Figure 4.1: Total Demand by Month and day of Week

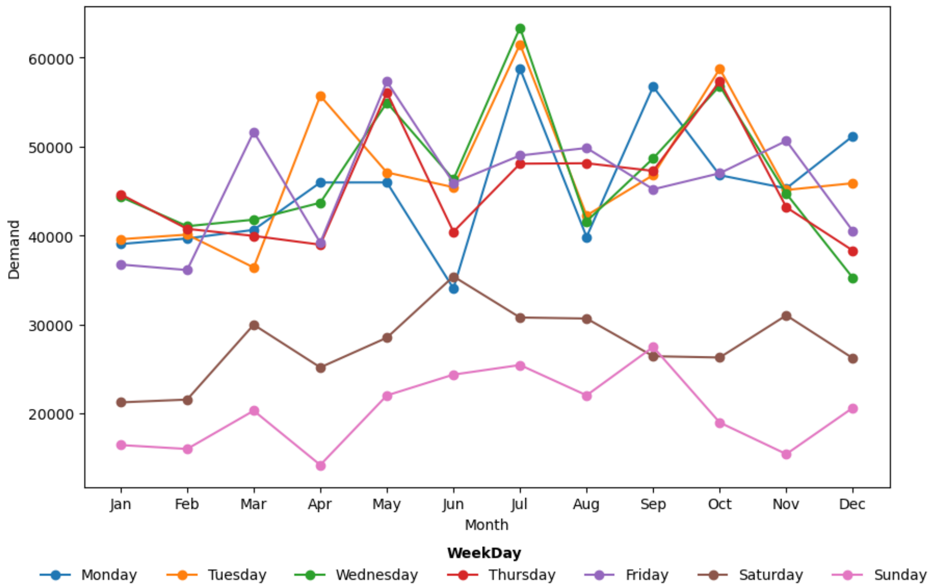


Figure 4.1 shows the distribution of the line 205 Andante validations by day of Week and by Month. The data seems to follow similar demand patterns for work days, which makes sense, considering these days have the same schedules and trips frequency. Therefore, for the model development, days were classified as workdays, Saturdays, or Sundays/holidays.

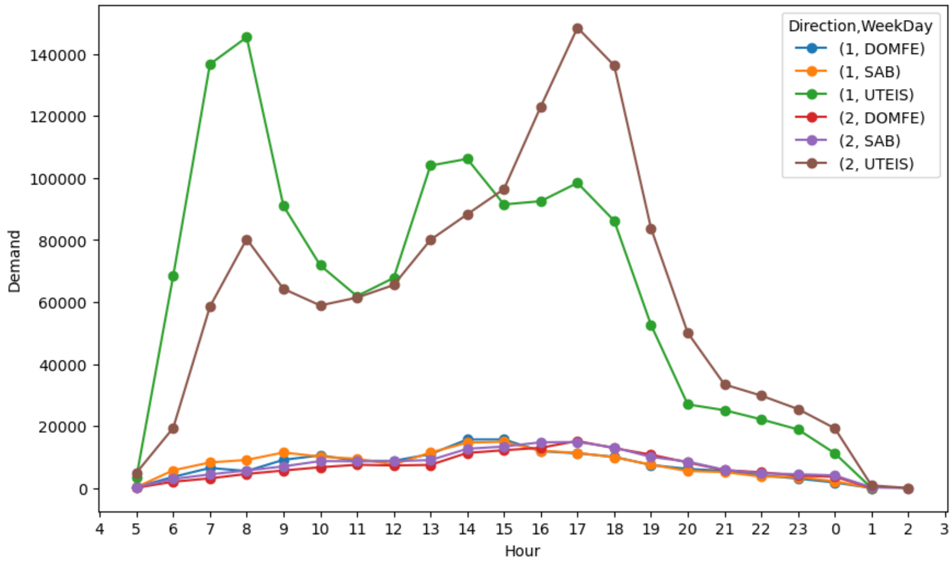
4.6.1 Daily Patterns Analysis

The distribution of passengers during the day, representing the daily travel patterns, is vital for load distribution. To better comprehend it, Figure 4.2 showcases the demand for Andante by hour, by direction and by weekday.

In Figure 4.2, only workweek trends can be analysed, since their demand is so much bigger than Saturdays and Sundays/holidays, that the later distributions appear linear in comparison.

For the workdays, the graph shows a distinct trend. According to the data, there are very few passenger journeys made in the early morning hours. As the day progresses, however, the number of journeys increases rapidly, reaching a first peak around 8:00, and then gradually declines. Between 12:00 and 13:00, around lunch time in Portugal, a second increase is noticeable. This is

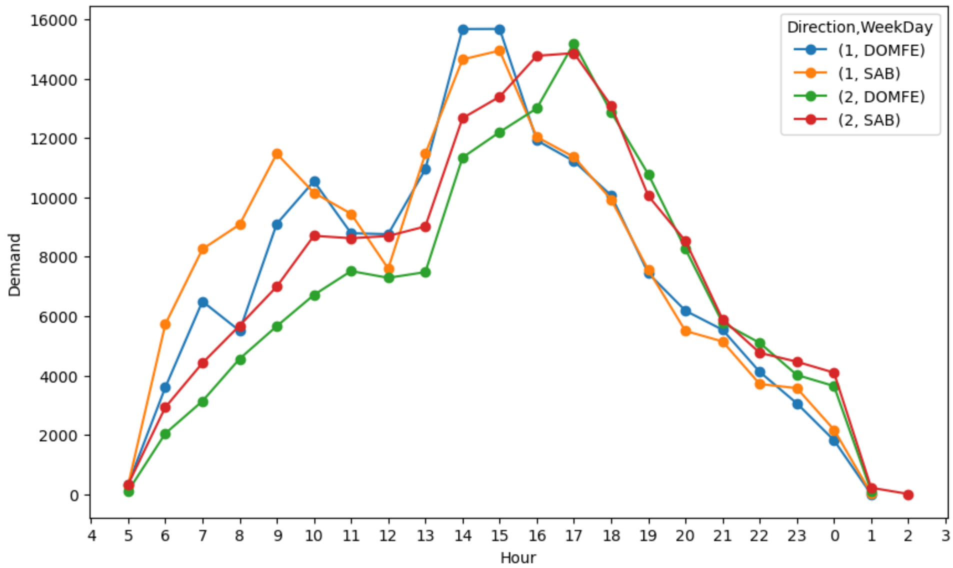
Figure 4.2: Andante Demand by Hour and day of Week



more prominent in direction 1, where it is followed by a subsequent decrease, while in direction 2 the number of validations continues to increase until the end of the afternoon. Around 17:00 and 18:00 a second peak is visible for direction 2, there is also an increase in direction 1, but it is not as prominent. After this time, the number of validations declines into the night.

The patterns for the opposite directions appear to be mirrored, with direction 1 showing a higher demand in the morning peak, and a lower demand in the afternoon peak and vice versa.

Figure 4.3: Andante Demand by Hour Weekend



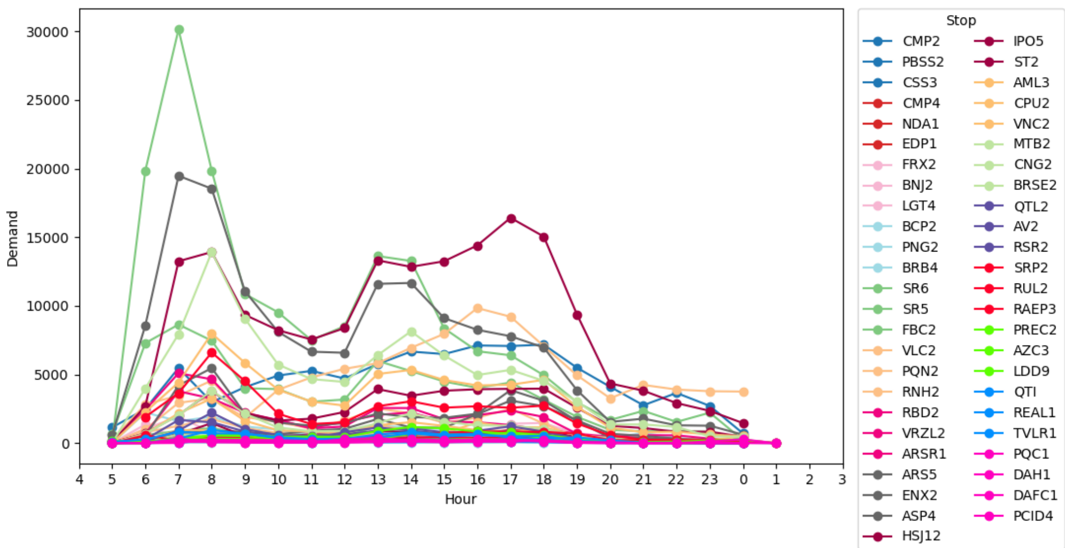
A second graph (Figure 4.3), showcasing the demand for Andante by hour and by Direction for Saturdays and Sundays/holidays was generated to analyse the patterns associated with these types of days.

Non work days seem to follow very similar patterns. There are very few passenger journeys made in the early morning hours. For direction 1 the number of journeys increases, reaching a first peak around 9:00 to 10:00, and then gradually declines. Around 12:00 there is a second, more prominent increase, that leads to a second peak around 14:00 to 15:00. Similarly, for direction 2, the number of journeys increases more steadily until 10:00 to 11:00, then appears stable until 13:00. At this time a second rapid increase happens, similar to the one from direction 1, leading to a peak in validations at 17:00.

Following their corresponding afternoon peaks, the number of validations for both directions decreases steadily until the end of the day.

In order to analyse the demand of stops according to hour of day, and visualise the stops correspondent to the peak hours, for each direction, the passenger demand according to stop throughout the day was plotted. Figure 4.4 relates to direction 1, while Figure 4.5 is referent to direction 2.

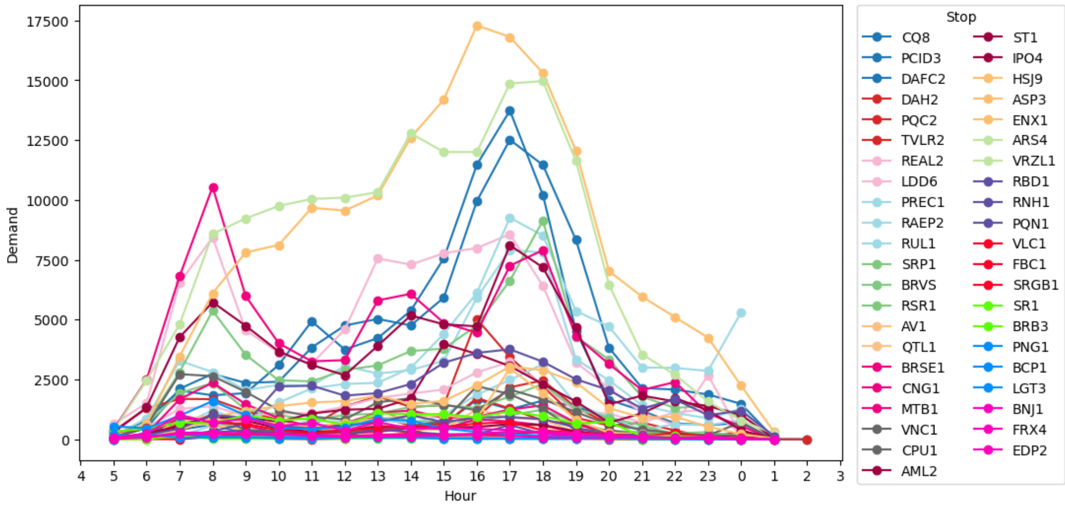
Figure 4.4: Direction 1: Demand by Stop and Hour



For direction 1 (Figure 4.4), the stops with the higher demand are consistently SR6, ARS5, HSJ12 and MTB2, and the demands fluctuate according to the direction 1 patterns. Most of these stops are part of the first half of the route, with the exception of MTB2. It is worth to note that the increase at 17:00 is mostly due to the increase in demand in HSJ12.

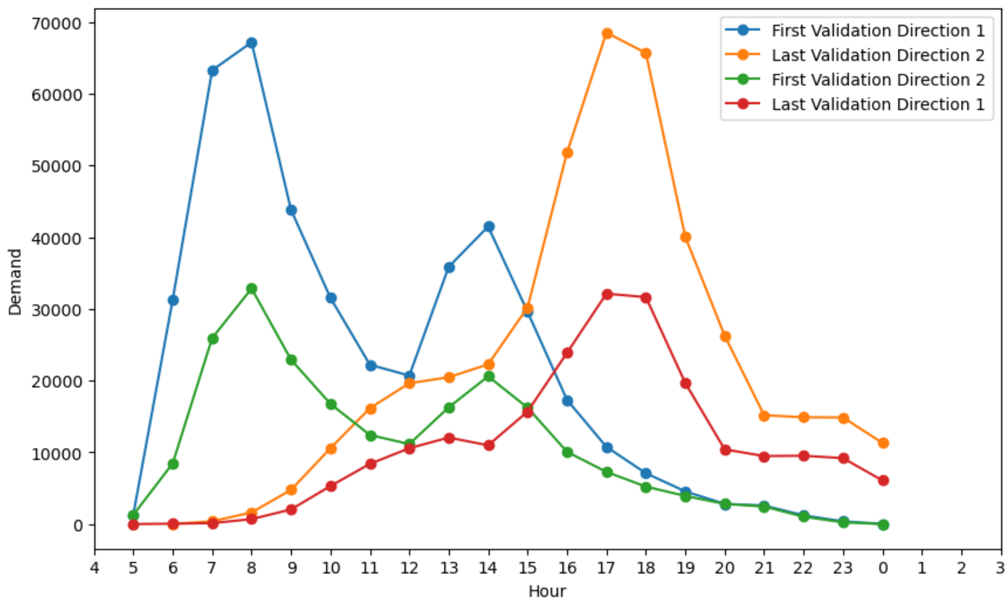
For direction 2 (Figure 4.5), the stops with the higher demand between 05:00 and 09:00, period of demand increase until the first peak, are MTB1, ARS4, LDD6, HSJ9, AML2 and SRP1. The stops with the higher demand between 15:00 and 19:00, period correspondent to the afternoon peak, are HSJ9, ARS4, PCID3, CQ8, LDD6 and RUL1. The demand for the stops AML2 and MTB1 also increases significantly during this period, following the direction 2 daily demand, however, PCID3, CQ8 and RUL1 are atypical, with a high demand exclusive to the afternoon. HSJ9 and ARS4 have a pretty high demand during the entirety of the day.

Figure 4.5: Direction 2: Demand by Stop and Hour



Contrary to direction 1, the stops that cause the morning peak are different from the ones that cause the afternoon peak.

Figure 4.6: Andante: First and Last Daily Validation Demand by Hour



An analysis of the daily patterns of passengers that traveled the same route in opposite directions in the same day was conducted. Passengers that fit this description at least once represented 41,73% of the total number of passengers.

Figure 4.6 presents the daily distributions of the time of the first and last daily validations by direction. Even though direction 1 has a superior demand during the morning period, and an inferior demand during the afternoon period and the case is the opposite for direction 2, the trend of the distribution of the first and last daily validations is pretty much the same, proportionately.

From this graph the following periods were identified:

- First daily validation peak: 07:00 to 8:00
- Last daily validation peak: 17:00 to 18:00
- First daily validation period: 05:00 to 11:00
- Last daily validation period: 15:00 to 00:00
- Undetermined period: 11:00 to 15:00

4.6.2 Stops Patterns Analysis

Analyzing demand patterns at different stops and directions can provide valuable information about passenger flow and help in understanding the occupancy dynamics of a bus line, for example which stops or segments of the route have higher passenger demand. This information can be useful in understanding the areas or time periods where the bus is likely to be more crowded or where there is a higher likelihood of passengers boarding or alighting.

Figure 4.7: Andante: Total Demand Distribution by Stop and Direction

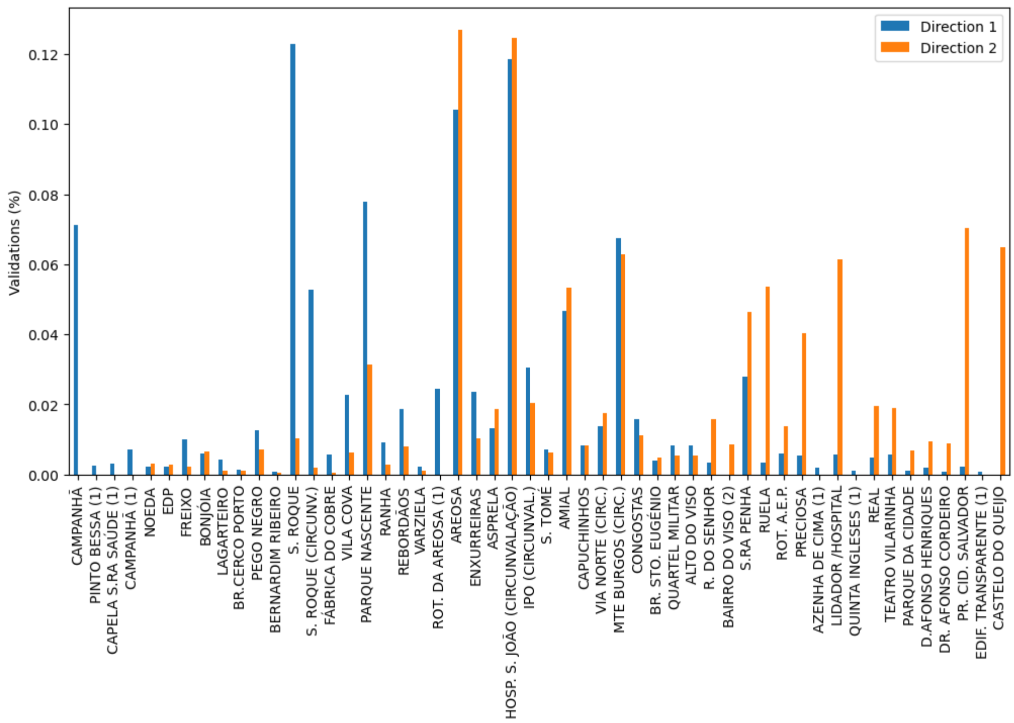
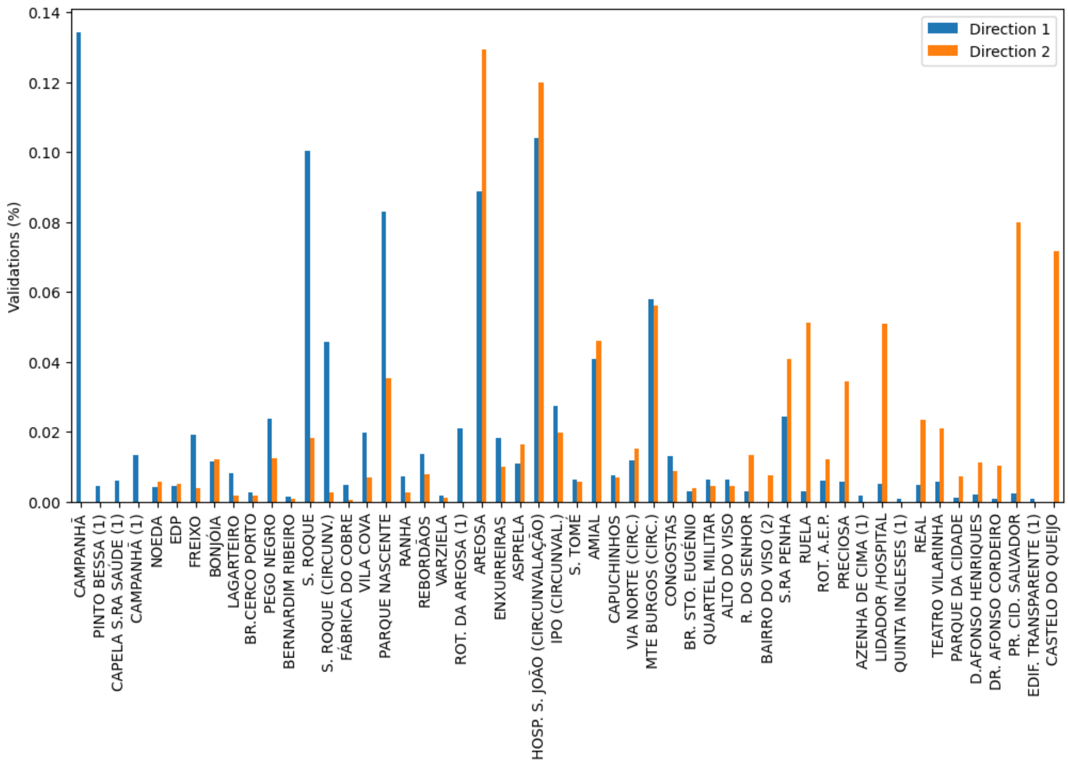


Figure 4.7 presents the distribution of andante validations by stop according to direction. This distribution is in percentages to better compare both directions, since they differ in total number of validations. Reading the graph from left to right the stops are ordered by the direction 1 route, from right to left it is possible to analyse the direction 2 route. The stops that are exclusive to one of the directions are signaled in the axis by the corresponding direction between parenthesis.

Since the bus line consists in a number of variants, that cover different sections of the total route, this chart has an unusually high demand on the middle stops. This makes it an accurate representation of the demand of the individual stops, however, it does not inform regarding demand in a trip. For this reason, additional graphs representing the different variants (described in 3.2) were generated.

Figure 4.8 represents the trips CMP-CQ and CQ-CMP, Figure 4.9 represents the trips SR-CQ and CQ-SR, and Figure 4.10 represents the trips SR-LDD and LDD-SR. Since the remaining routes do not have a corresponding variant in the opposite direction, they were not analysed.

Figure 4.8: CMP-CQ: Demand Distribution by Stop and Direction



All of the analysed variants show a consistent higher demand in stops Areosa, Hospital de S.João (Circunvalação), Amial, Mte Burgos (Circ.) and S.ra Penha, in both directions.

With the exception of these stops with higher traffic, the trend for the route seems to be that the demand is superior in the beginning of the line, and decreases until the end of the line, where it is relatively low. Another observation is that for the stops in the middle of the route, the distribution of both directions is consistently quite similar.

In order to evaluate the pertinence of the analysis of the Anda dataset as representative of the Andante dataset, a comparison of their respective percentages of entries by stops was conducted. Table 4.7 shows the percentage differences between the Anda and Andante dataset by stop. Stops with a percentage difference inferior to 1% are coloured green, if this value is between 1% and 3%, it is coloured yellow, in the case the difference is superior to 3% it is coloured red.

Figure 4.9: SR-CQ: Demand Distribution by Stop and Direction

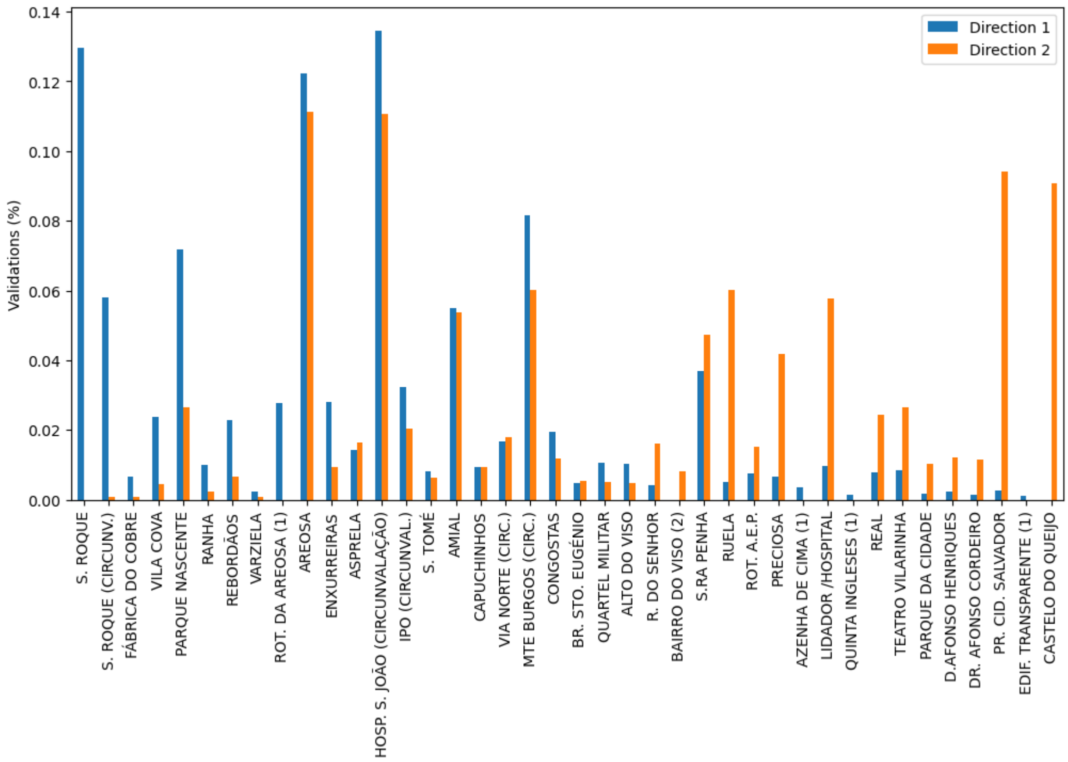
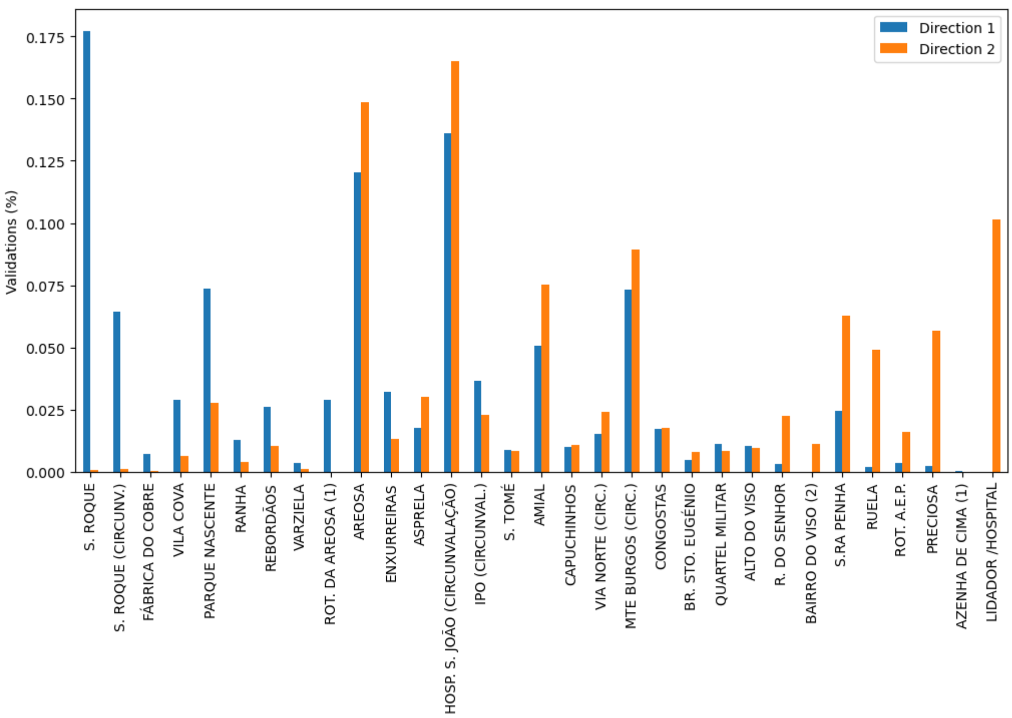


Figure 4.10: SR-LDD: Demand Distribution by Stop and Direction



The second Campanhã stop (CMP4), exclusive to direction 1 could not be identified in the Anda dataset, as it is a homonym of the first stop of the route (CMP2), and the Anda dataset only contains stop names, that do not act as an ID in this particular case, as previously mentioned. Since CMP4 had a consistently higher demand in the Andante dataset, it was assumed that Campanhã values in the Anda dataset referred to it.

For direction 1, containing 51 stops, the percentage difference between Anda and Andante is superior to 3% for 4 stops, with a maximum value of 6,20% and is a value between 1% and 3% for 7 stops. For direction 2, containing 45 stops, the percentage difference between Anda and Andante is superior to 3% for 5 stops, with a maximum value of 7,16% and is a value between 1% and 3% for 11 stops.

Overall, the datasets are similar, and it was considered that the Anda from 2019 was representative of the Andante dataset from 2019, for line 205.

Zone	Stop Name	Direction 1	Direction 2
PRT1	Campanha	2.88%	0.00%
	Pinto Bessa (1)	0.15%	-
	Capela S.Ra Saude (1)	0.22%	-
	Campanha (1)	-	-
	Noeda	0.06%	0.28%
	Edp	0.09%	0.21%
PRT3	Freixo	0.76%	0.16%
	Bonjoia	0.39%	0.40%
	Lagarteiro	0.32%	0.10%
	Br.Cerco Porto	0.07%	0.11%
	Pego Negro	0.55%	0.48%
	Bernardim Ribeiro	0.01%	0.05%
	S. Roque	6.20%	1.70%
	S. Roque (Circunv.)	3.42%	0.04%
	Fabrica do Cobre	0.38%	0.01%
	Vila Cova	0.44%	0.20%
	Parque Nascente	2.48%	1.21%
	Ranha	0.76%	0.08%
	Rebordaos	1.68%	0.62%
	Varziela	0.13%	0.02%
	Rot. da Areosa (1)	0.36%	-
	Areosa	1.04%	7.16%
	Enxurreiras	2.32%	0.33%
	Asprela	2.74%	0.90%
	Hosp. S. Joao (Circunvalacao)	3.71%	2.97%
	Ipo (Circunval.)	2.20%	0.04%
	S. Tome	0.61%	0.21%
	Amial	3.10%	4.35%
PRT2	Capuchinhos	0.77%	0.23%
	Via Norte (Circ.)	1.93%	1.18%
	Mte Burgos (Circ.)	0.10%	1.72%
	Congostas	0.05%	1.08%
	Br. Sto. Eugenio	0.72%	0.12%
	Quartel Militar	0.29%	0.94%
	Alto do Viso	0.19%	0.90%
	R. do Senhor	0.23%	0.69%
	Bairro do Viso (2)	-	0.41%
	S.Ra Penha	0.40%	1.03%
	Ruela	0.29%	3.41%
	Rot. A.E.P.	0.05%	1.70%
	Preciosa	0.33%	0.74%
	Azenha de Cima (1)	0.10%	-
	Lidador /Hospital	0.42%	3.32%
	Quinta Inglezes (1)	0.04%	-
	Real	0.06%	0.02%
	Teatro Vilarinha	0.18%	1.10%
	Parque da Cidade	0.09%	0.91%
	D.Afonso Henriques	0.16%	2.50%
	Dr. Afonso Cordeiro	0.03%	2.47%
	Pr. Cid. Salvador	0.17%	1.88%
	Edif. Transparente (1)	0.04%	-
	Castelo do Queijo	0.00%	3.94%

Table 4.7: Percentage difference between Andante and Anda Validation distributions by stop

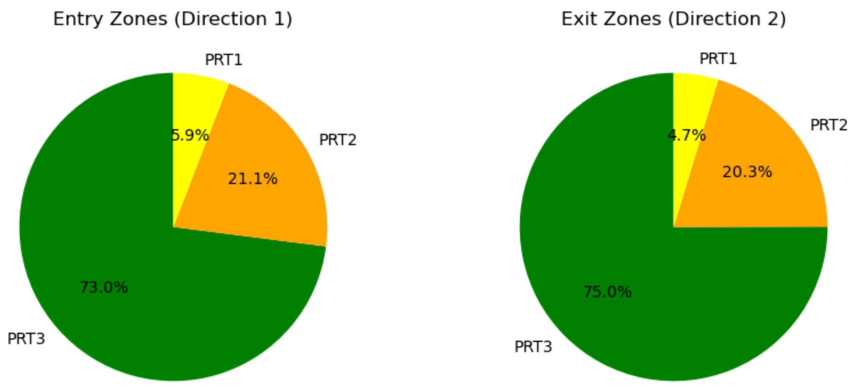
4.6.3 Entries and Exits Patterns Analysis

Analyzing a representative dataset with trip's entry and exit stops information can be can provide valuable insights and contribute to accurate occupancy estimation.

The Anda dataset serves as a reliable source of ground truth data for a subset of the bus line's trips. By capturing actual entry and exit information from passengers, this dataset provides a reference point for understanding the occupancy levels at different stages of the journey.

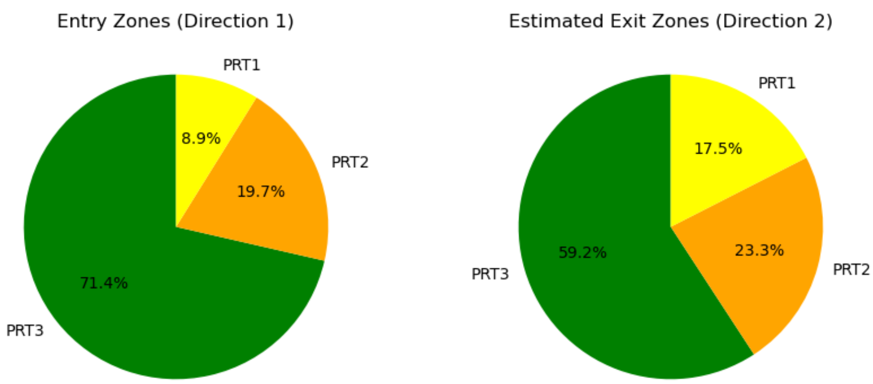
Since the only information regarding exits present in Andante is the exit zone, a comparative analysis of the entries and exits zones from both datasets was conducted.

Figure 4.11: Anda: Entry Zones Direction 1 and Exit Zones Direction 2



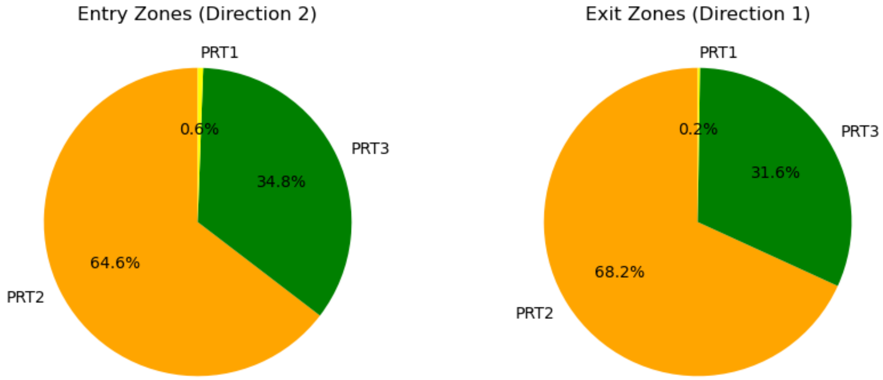
For Anda, the distribution of entry zones from direction 1 was close to identical to the distributions of exit zones from direction 2, with a maximum difference of 2%, as seen in Figure 4.11.

Figure 4.12: Andante: Entry Zones Direction 1 and Estimated Exit Zones Direction 2



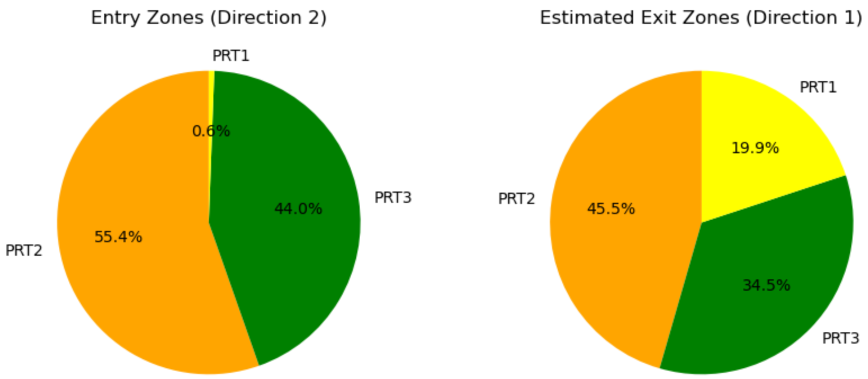
However, even though the Andante entry zones for direction 1 were similar to the Anda ones, this pattern was not observable in the Andante data, as seen in Figure 4.12. Comparing Andante's direction 1 entry zones and direction 2 exit zones, the distributions varied considerably more, with a maximum difference of 12,2%.

Figure 4.13: Anda: Entry Zones Direction 2 and Exit Zones Direction 1



Comparing the Anda's direction 2 distribution of entry zones to the direction 1 exit zones, the same pattern was observable, with a maximum discrepancy of 3,6%, as seen in Figure 4.13.

Figure 4.14: Andante: Entry Zones Direction 2 and Estimated Exit Zones Direction 2



For the Andante dataset, where direction 2 entry zones were, once again, comparable to the ones from Anda, the distributions were quite different. As seen in Figure 4.14, the direction 2 entry zones and direction 1 exit zones, had a difference as high as 19,3%.

This analysis further confirmed that the *'ExitZone_RR'* column from the Andante dataframe does not have accurate information regarding the alighting zones from the passenger trip corresponding to the validations.

Taking into consideration how similar the entries and exits from opposite directions appeared, an analysis at stop level was conducted.

Figure 4.15 compares the demand of direction 1's entry stops, with direction 2's exit stops.

Figure 4.16 compares the demand of direction 2 entry stops, with the demand for direction 1 exit stops.

These comparisons are made in percentage to account for the difference between total number of validations by direction. The distributions are remarkably similar, showing a relation between

Figure 4.15: Anda: Direction 1 Entries and Direction 2 Exits by Stop

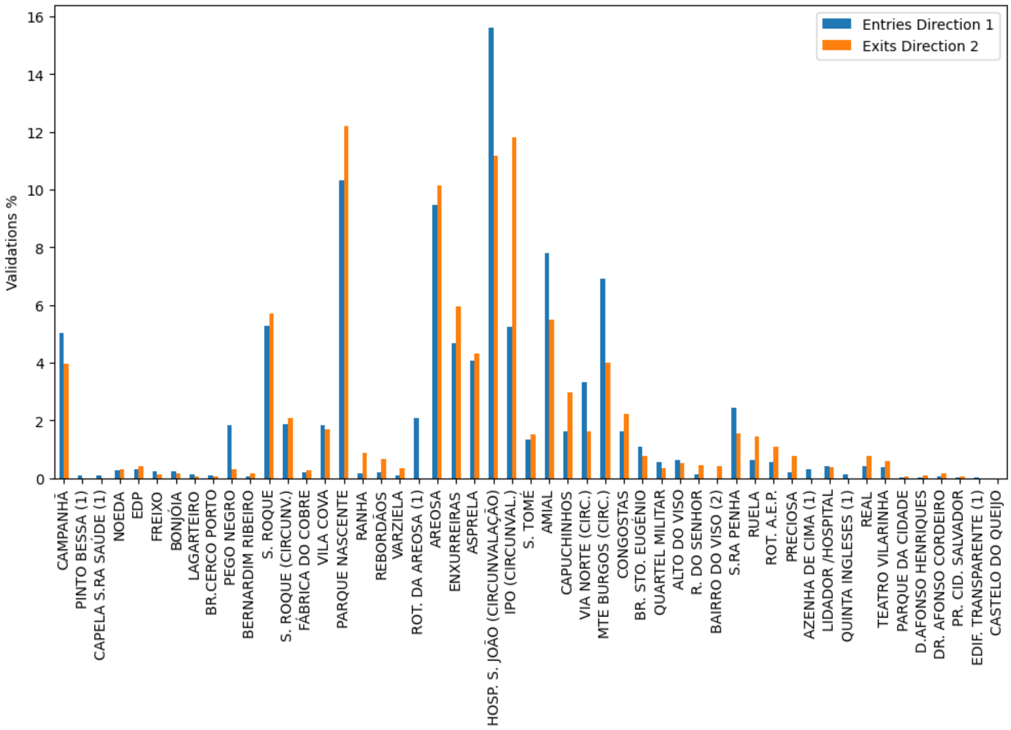
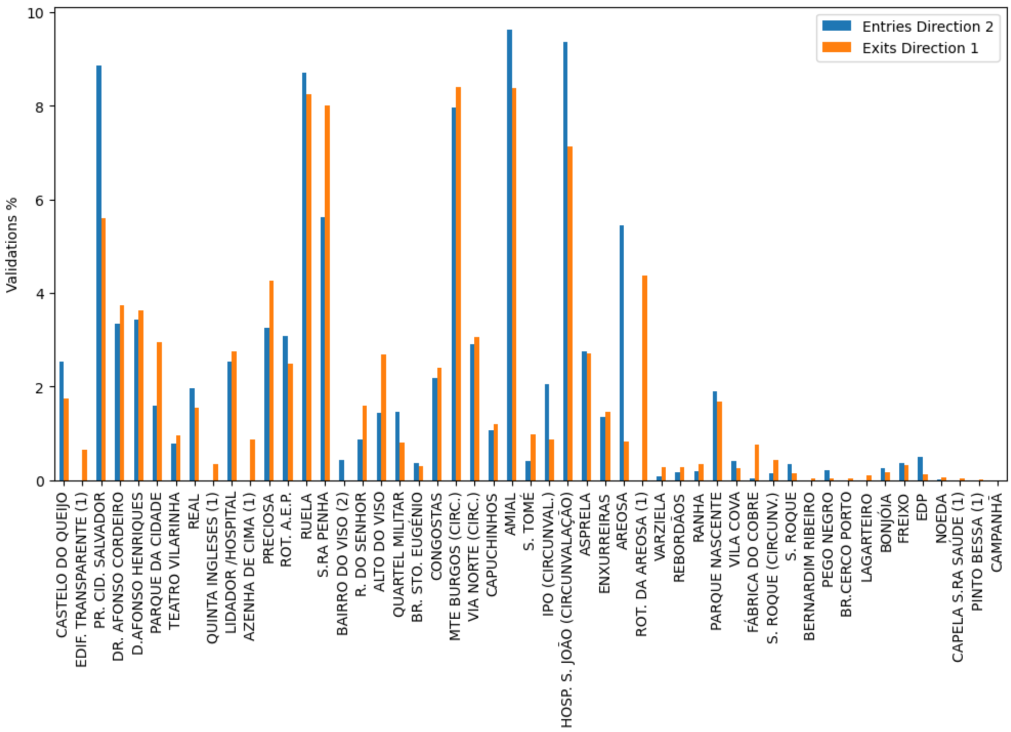


Figure 4.16: Anda: Direction 2 Entries and Direction 1 Exits by Stop



the demand of stops for boarding and alighting.

For the stops exclusive to one of the directions, it is not possible to make this comparison. However, it can be assumed that passengers who board at a certain stop in one direction, alight at the closest stop when traveling in the opposite direction. This can be observed in Figure 4.16, by comparing the Areosa and Rot. da Areosa (1) stops.

With this method, Table 4.8 was generated. This table shows the percentage difference between opposite direction's entries and exits, comparing stops directly when possible. For stops exclusive to one direction, the comparison is made to the set of corresponding stops, defined previously, making a total of 39 stops and 5 sets of stops.

Percentage differences inferior to 1% are coloured green, between 1% and 3% are coloured yellow, and differences superior to 3% are coloured red.

Entries from direction 1 are extremely similar to exits from direction 2, with only 2 cases where the variance is superior to 3% and 9 cases where it is a value between 1% and 3%.

Comparing the entries from direction 2 to the exits from direction 1, the same pattern is observable, with only 8 cases where the variance is a value between 1% and 3%.

Overall, this analysis shows a correlation between entry and exit stops in opposing directions. Because of the high degree of similarity in this correlation, it is possible to assume that the occupancy patterns of the bus line can be estimated by examining the entry data alone. Additionally, this consistency supports the assumption of passenger behavior used in this scenario.

In conclusion, the discovery of a consistent pattern between entry and exit stops in opposite directions supports the development of a reliable model for estimating the number of passengers on board throughout the bus line, even in the absence of exit data.

4.7 Bus Line Occupancy Estimation Model

This section describes the development of the bus line occupancy estimation model. The goal of this model is to estimate the number of passengers on board at various stops along the bus line. Since the actual exit information is not available, the model makes use of factors deemed as relevant during the data analysis process to infer the number of exits at each stop, subsequently calculating the estimated occupancy at each stop. These factors include the time of day, day of the week, and the characteristics of different stops and directions.

The occupancy estimation model relies on the entry data collected from the bus line. This data includes the number of passengers boarding at each stop and the corresponding timestamps, as described in Trip Concept Definition.

The estimation model was used the line 205 data and used some information exclusive to this line, regarding the route and variations only. This information is available in the STCP website for all bus lines, meaning that the methodology developed could be applied to any bus line.

Zone	Stop Name	Entries 1 - Exits 2	Entries 2 - Exits 1
	Campanhã		
	Pinto Bessa (1)		
	Capela S.Ra Saúde (1)	1.26%	0.06%
PRT1	Campanhã (1)		
	Noeda	0.03%	0.04%
	Edp	0.09%	0.38%
	Freixo	0.13%	0.05%
	Bonjóia	0.05%	0.08%
	Lagarteiro	0.07%	0.09%
	Br.Cerco Porto	0.04%	0.04%
	Pego Negro	1.51%	0.18%
	Bernardim Ribeiro	0.10%	0.04%
	S. Roque	0.42%	0.20%
	S. Roque (Circunv.)	0.20%	0.30%
	Fábrica do Cobre	0.06%	0.72%
	Vila Cova	0.16%	0.16%
	Parque Nascente	1.90%	0.21%
PRT3	Ranha	0.73%	0.15%
	Rebordãos	0.48%	0.11%
	Varziela	0.24%	0.19%
	Rot. da Areosa (1)		
	Areosa	1.41%	0.23%
	Enxurreiras	1.28%	0.11%
	Asprela	0.27%	0.05%
	Hosp. S. João (Circunvalação)	4.46%	2.24%
	Ipo (Circunval.)	6.58%	1.18%
	S. Tomé	0.20%	0.57%
	Amial	2.31%	1.27%
	Capuchinhos	1.36%	0.14%
	Via Norte (Circ.)	1.69%	0.14%
	Mte Burgos (Circ.)	2.89%	0.45%
	Congostas	0.60%	0.22%
	Br. Sto. Eugenio	0.34%	0.07%
	Quartel Militar	0.20%	0.67%
	Alto do Viso	0.10%	1.26%
	R. do Senhor	0.34%	0.73%
	Bairro do Viso (2)		
	S.Ra Penha	0.47%	1.95%
	Ruela	0.82%	0.48%
	Rot. A.E.P.	0.53%	0.59%
PRT2	Preciosa	0.57%	1.00%
	Azenha de Cima (1)		
	Lidador /Hospital	0.46%	1.44%
	Quinta Inglezes (1)		
	Real	0.35%	0.40%
	Teatro Vilarinha	0.23%	0.17%
	Parque da Cidade	0.04%	1.37%
	D.Afonso Henriques	0.08%	0.18%
	Dr. Afonso Cordeiro	0.10%	0.39%
	Pr. Cid. Salvador		
	Edif. Transparente (1)	0.04%	2.61%
	Castelo do Queijo	0.00%	0.80%

Table 4.8: Anda: Percentage difference between opposite direction's entries and exits

4.7.1 Assumptions and Insights from Data Analysis

To develop the occupancy estimation model, extensive data analysis and pattern identification were performed on the entry data. This analysis revealed important insights and allowed the formulation of key assumptions about passenger behavior and boarding patterns.

Based on the data analysis, the following assumptions were made:

Passenger Retention:

It is assumed that passengers who board at a particular stop will stay on the bus for at least two stops. This assumption aims to ensure that the number of estimated descents is never superior to the number of passengers in the bus, considering a conservative minimum trip relatively to travel patterns.

Travel Patterns:

By studying the entry data, patterns related to peak boarding times, variations in boarding behavior on different days of the week, and the influence of specific stop locations on passenger volumes were identified. To estimate the number of exits at each stop, the model leverages these observed travel patterns and the entries from the opposite direction. The model assumes that the distribution of exits in one direction is similar to the distribution of entries in the opposite direction for corresponding trips, defined based on the following factors:

- **Same Seasonality**: The model compares the exits in one direction with the entries in the opposite direction that occurred during the same season. For this purpose, the seasons defined by STCP, when schedule changes occur, were considered. In 2019 these consisted in: *Summer*, starting with the Summer school vacations, from 15/06/2019 to 26/07/2019; *August*, from 27/07/2019 to 01/09/2019; *Holidays*, corresponding to the Christmas school vacations from 19/12/2019 to 02/01/2020; and *Regular* season, the rest of the year.

- **Same Weekday**: The model considers the exits in one direction and the entries in the opposite direction that happened on the same weekday. The weekdays are categorized as either working days or weekends/holidays.

- **Same Variant**: The model ensures that the distributions being compared are from the same variant. Variants that follow the same route in both directions are corresponded directly (see Table 3.2). This is the case for CMP-CQ and CQ-CMP, complete trips, SR-CQ and CQ-SR, considered sub-trips 1, and SR-LDD and LDD-SR, sub-trips 2. The variants LDD-CMP and HSJ-CMP do not have corresponding variants in the opposite direction. Because of this they are matched to the shortest route that contains all of their stops, (either Complete, Var1 or Var2) in this case complete trips. The remaining variants MTB-CQ and CQ-MTB are matched to sub-trips 1, and MTB-LDD and LDD-MTB are considered sub-trips 2. This was decided because even though these variants follow the same route in opposite directions, very few trips were identified in the Andante dataset, therefore there is not enough information regarding these variants to follow the established method (see Table 4.6.

- **Corresponding Period**: The model divides the day into specific periods, according to the insights obtained from the Daily Patterns Analysis. The First Peak (7:00 to 8:00) is corresponded to Last Peak (17:00 to 18:00) and vice-versa, First Val (5:00 to 11:00) is corresponded to Last Val (15:00 to 00:00) and vice versa. For trips occurring between 11:00 and 15:00 (Undefined period), the model considers their distribution relative to the entire day.

By considering these factors, the model infers the number of exits at each stop based on the observed entries from the opposite direction during corresponding trips.

Corresponding Stops:

As previously explained in Entries and Exits Patterns Analysis, the stops from opposite directions are corresponded based on distance, according to the passenger behaviour assumption that boardings at a certain stop in one direction will alight at the closest stop when traveling in the opposite direction. This means that stops are corresponded directly when possible, and for stops exclusive to one direction, the comparison is made to the set of corresponding stops. For the latter case, the model assumes that the distribution of exits follows the same distribution of entries for the corresponding set of stops. This assumption ensures a consistent estimation of occupancy levels throughout the route.

Terminal Descent:

It is assumed that all passengers who leave the bus when it reaches the terminal stop.

4.7.2 Occupancy Estimation Process

The occupancy estimation model follows a straightforward process:

Input: The model receives the entry data for a particular bus trip, which includes the number of passengers boarding at each stop and the corresponding timestamps, as well as the information regarding all parameters related to the assumptions.

Opposite entries distributions (*dist*): The model retrieves the entries distributions from the correspondent trips, as explained in the Travel Patterns assumptions, and adapts them according to the current trip using the following process:

1. **Exclusive stops information:** Since this distribution information is regarding the opposite direction, it adds the information regarding the stops exclusive to the current direction, using the method explained in the Corresponding Stops assumptions.

2. **Flattening transformation:** Since the distributions correspond to very high numbers of validations, exponentially bigger than actual trips' occupancy, a flattening transformation of $dist^{0,75}$ is applied to the vector.

3. **Percentage conversion:** Then, the distributions are converted to percentages. In the case that the variant from the trip in question does not directly match the correspondent trips, the non-relevant stops are excluded from the vector prior to this conversion. As an example, a

LDD-CMP trip is classified as a *Complete* due to lack of a better opposite direction correspondence, meaning the distributions include 7 stations from CQ-LDD that are not a part of the trip route. These are irrelevant, and are therefore excluded.

Initial Occupancy (*occ*): The exit values are initiated as 0, and the model calculates the initial occupancy at each stop, using the values of the entries accumulated subtracted by the values of the exits accumulated.

$$\mathrm{occ}_i = \sum_{j=1}^{i} entries_j - \sum_{j=1}^{i} exits_j$$

Iterative Estimation: Starting from the first stop, the model iteratively estimates the number of exits. The exits are assumed to follow the same distribution from the vector previously calculated. Therefore, for each stop, the initial occupancy is retrieved, the percentage from the distributions is retrieved, and the exits are assumed to be the corresponding percentage of the initial occupancy.

$$\mathrm{exits}_i = \mathrm{occ}_i \times \mathrm{dist}_i$$

Before continuing the cycle to the next stop, the *occ* is updated using the same formula presented in the previous step.

Exits Adjustment: Using the explained method, the estimated exits do not equate the number of passengers in the trip. However, it is assumed that no passengers remain in the bus after reaching the terminal, for this to be true, the total number of entries needs to be equal to the total number of exits. Therefore, using cross-multiplication, the estimated exits are adjusted accordingly.

$$\mathrm{nr\_exits} = \sum_{i=1}^{n} \mathrm{exits}_i$$

$$\mathrm{exits}_i = \frac{\mathrm{nr\_passengers} \times \mathrm{exits}_i}{\mathrm{nr\_exits}}$$

Exits Rounding: Since the estimation being made is regarding passengers, it needs to correspond to whole numbers. Therefore, the estimated exits are rounded, ensuring once again, that the total sum equates the total entries.

Access Passenger Retention Assumption: The exits estimations are evaluated to check whether the restriction that passengers traveled at least two stops was met. If it was not, the vector of maximum number of exits by station is adjusted so that its sum is equal to the sum of estimated exits calculated by the *Iterative Estimation* step. This estimation is then performed again, checking in each station weather the number of exits estimated surpasses the allowed, in which case it is set to the allowed. Subsequently the *Exits Adjustment* and *Exits Rounding* are re-done.

Final Occupancy: Finally, using the entries values and the exit estimation values, the occupancy at each stop is estimated and returned.

4.7.3 Methodology Pipeline

The complete pipeline followed during this study (excluding the extensive data analysis) is described in figure 4.17. Each block corresponds to a very distinct phase, developed in separate from the others. The inputs and outputs from every one of this processes can be consulted, as well as the order in which they were followed. The only block in this pipeline that includes manually added parameters (not directly extracted from input files) is the Line Data Preparation, where some processing of the intermediate terminals is done. Because of this, this pipeline is very easily applied to other lines.

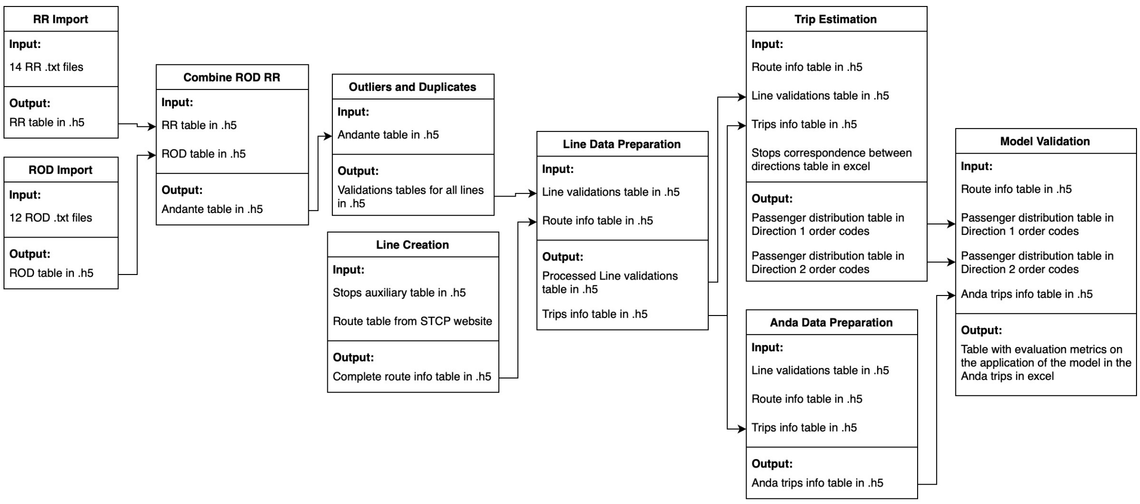


Figure 4.17: Methodology Pipeline

Chapter 5

Results and Discussion

The main objective of this chapter is to analyze and discuss the outcomes of the applied methodology, described in Chapter 4, in order to gain insights and draw meaningful conclusions.

This is done by presenting the evaluation method, and interpreting the results of the model validation process, in order to evaluate the effectiveness of the model in estimating passenger occupancy based on entry data alone.

As the approach used for occupancy estimation was based on the estimation of exits by stop, followed by calculation of the occupancy, the exits estimation was evaluated in addition to the occupancy. This was done because it can provide important information regarding the errors that originate the inaccuracies in the model.

5.1 Evaluation Method

To evaluate how well the model predicted the actual occupancy of the bus line, it was applied to the Anda dataset, that was considered representative of Andante, ensuring consistency and relevance in the comparison.

5.1.1 Data Selection

The Anda dataset had a maximum value of trips per day of around 30, and a maximum values of passengers in a single trip of 5. Because of this, in order to simulate complete bus trips information using the dataset, the validations were grouped by factors that showed similar travel patterns, including month, day of week, period in the day and variant. This resulted in the total of 637 trips. For the validation of the method, the 293 trips with a number of passengers superior to 20 were selected.

5.1.2 Model Evaluation

To generate the occupancy predictions, the developed model was applied to the entry data of the selected Anda trips.

Then, the predicted occupancy values were compared with the real ones, calculated using the actual exit information from Anda. This allowed for a direct evaluation of the model's accuracy, using various appropriate metrics.

5.1.3 Evaluation Metrics

Several evaluation measures were used to assess the performance of the occupancy estimation model, these metrics provide quantitative measures to evaluate the accuracy and reliability of the predictions.

Mean Absolute Error (MAE): Average absolute difference between the predicted and actual occupancy values. It measures the average magnitude of the errors, regardless of their direction. Lower values indicate better accuracy.

$$MAE = \sum_{i=1}^{n} |x_i - y_i|$$

n: Nr of observations, x_i: Actual value at i, y_i: Predicted value at i

Root Mean Square Error (RMSE): It is the square root of the MSE and similarly provides a measure of the average magnitude of the prediction errors. This measure penalizes larger errors more heavily compared to MAE. Lower values indicate better accuracy.

$$RMSE = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

n: Nr of observations, x_i: Actual value at i, y_i: Predicted value at i

Normalised Mean Absolute Error (NMAE): MAE normalised by the range of occupancy values, making it a value between 0 and 1.

$$NMAE = \frac{MAE}{\max(x) - \min(x)}$$

x: Actual values

Normalised Root Mean Squared Error (NRMSE): RMSE normalised by the range of occupancy values, making it a value between 0 and 1.

$$NRMSE = \frac{RMSE}{\max(x) - \min(x)}$$

x: Actual values

Explained Variance Score (EVS): This metric calculates the proportion of variance in the target variable that is explained by the model. It provides a measure of how well the model captures the underlying patterns and variations in the data. Explained variance score ranges from

0 to 1, with 1 indicating a perfect fit.

$$VAR(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$VAR(x)$: Variance, n: Nr of observations, x_i: Actual value at i, $\bar{x}$: Mean value of x

$$ESV = \frac{VAR(y-x)}{VAR(x)}$$

$VAR(x)$: Variance, x: Actual values, y: Predicted Values

These metrics facilitate the quantitative assessment of the accuracy and performance of the model in predicting the number of passengers at each stop. Through the examination of these metrics, insights can be gained into the effectiveness of the model and areas for improvement can be identified.

In the subsequent sections, the results of the model evaluation using these metrics will be presented, and the implications for evaluating the performance and reliability of the occupancy estimation model will be discussed.

5.2 Results

This section provides an analysis of the overall results obtained from the occupancy estimation model. Additionally, it explores the results based on specific characteristics of the trips, including outward vs return trips, weekday vs weekend trips, daily period, and line variant.

As previously established, only trips with a minimum number of 20 passengers were used for the model evaluation. This decision was primarily made since public transport buses have a capacity of more than 20 seated passengers, making the previously mentioned trips irrelevant, both in terms of fleet planning and in terms of passenger comfort. Additionally, due to the low occupancy values from these trips, the estimation errors associated were significantly lower, however, so was the coefficient of determination, indicating a lower performing model.

	Occupation				
	MAE	**MAN**	**RMSE**	**RMSN**	**EVS**
AVR	2,555	0,113	3,526	0,155	0,803
STD	1,651	0,057	2,246	0,071	0,176
MIN	0,633	0,037	0,983	0,055	-0,447
MAX	10,711	0,349	14,579	0,434	0,982

Table 5.1: Overall Occupation Estimation Results

The model results for occupancy estimation at each stop, are summarized in Table 5.1, and the results of the auxiliary exits estimation are in table 5.2. Again, the model was applied in 293 trips, the values of the average (AVR), the standard deviation (STD), as well as the minimum (MIN) and maximum (MAX) for each metric are highlighted in the table.

The occupation MAE ranges from 0,633 to 10,711, and there is an error of approximately 2,555 passengers in the occupancy estimation at each stop, which is considerably low for the context. The occupancy RMSE ranges from 0,983 to 34,921, and is around 3,526, indicating the average magnitude of the estimation errors. Figure 5.1 shows the distribution of all of the occupation metrics in box plots.

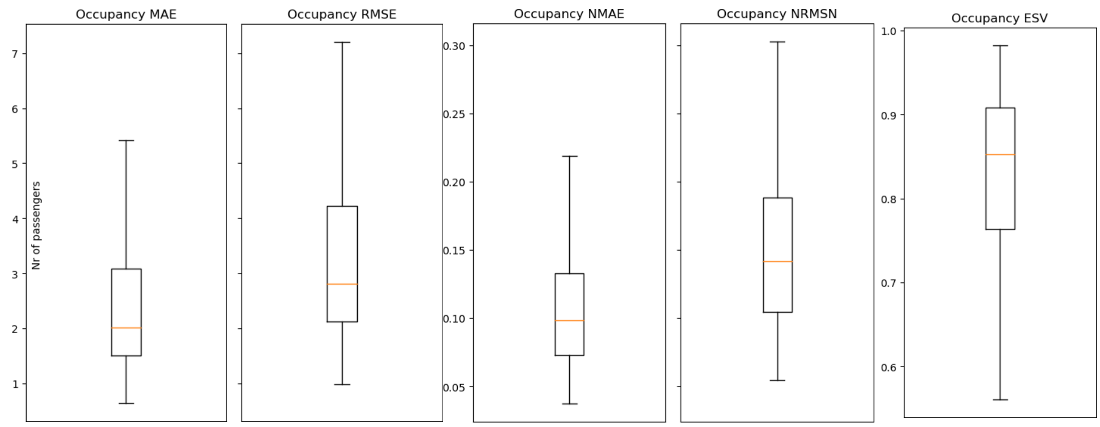


Figure 5.1: Occupation MAE, NMAE, RMSE, NRMSE and ESV Box plots

The EVS values indicate that the estimated occupancy accounts for approximately 80.3% of the variability in the actual occupancy. The values range from -0,447, which can be interpreted as the worst performing trip to 0,982, which can be interpreted as the best performing trip. A negative EVS indicates that the model's estimated values explain less variance than using a baseline model, calculated by the simple average of the occupancy, indicating a model poorly fitted to the data.

Figure 5.2 plots the values of the estimated and actual occupancy in each stop, for the worst performing trip. This is a CMP-CQ trip, with 32 passengers, taken in a working day of the regular season during the First Val period. Out of all evaluated trips, this one was the only one with a negative ESV value. The trip occurred in the First Val period of a working day, during regular season.

In this particular case, even though the amount of passengers exiting during the trip is poorly estimated, the exits' function still appears similar, just less accentuated. This trip in particular in the stop HSJ12 had an occupancy drop from 12 to 7 passengers, which corresponds to 42%. This would always be a hard value to estimate because of how unusual it is, even if the model was not constructed using distributions.

Additionally, it is worth noting that even though the peak is wrongly estimated, and at the wrong stop, the real number of maximum simultaneous passengers is 14 instead 18, which does not constitute a major error in terms of transport planing or passenger comfort.

The best performing trip can be consulted in Figure 5.3.

Figure 5.2: Worst performing trip: Actual and Predicted Occupancy

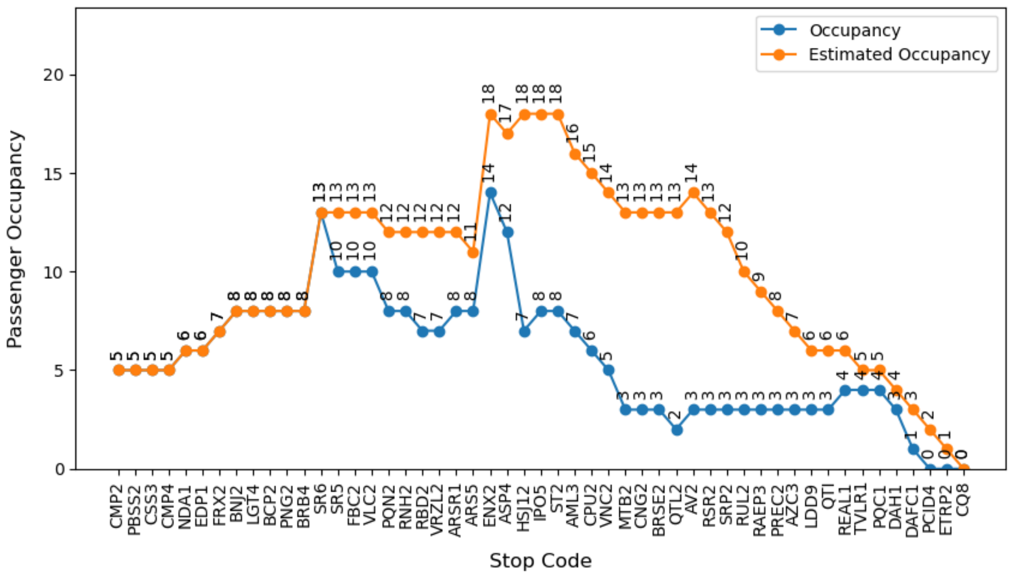
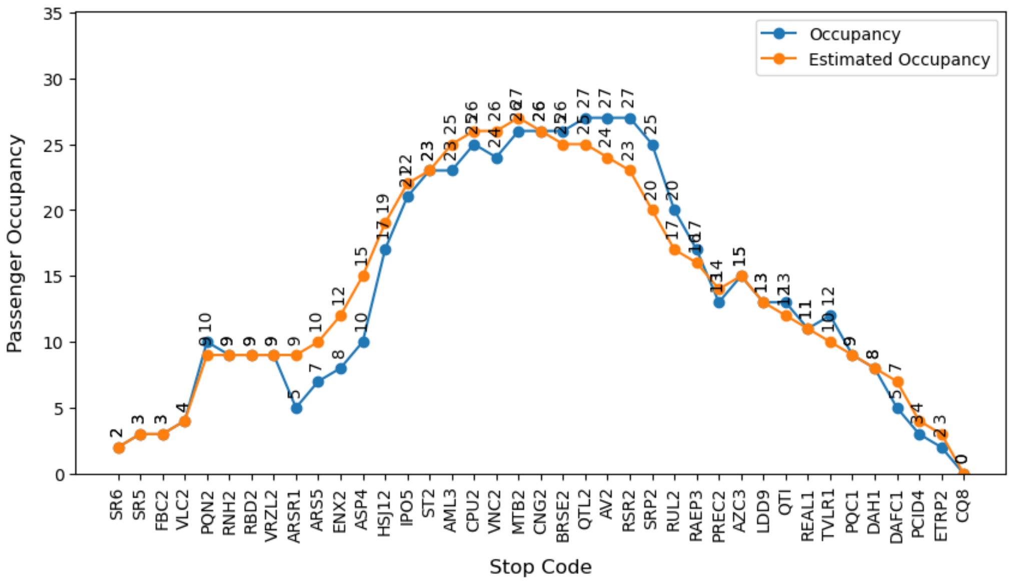


Figure 5.3: Best performing trip: Actual and Predicted Occupancy



The trip occurred in the undefined period of a working day, during regular season. In this particular case, the trip seems to follow a normal distribution, in contrast to the previous one, that appeared very irregular.

Overall, the results show that the model achieves reasonably good performance in estimating both occupancy and exits, as indicated by the low error metrics for the estimations in both cases.

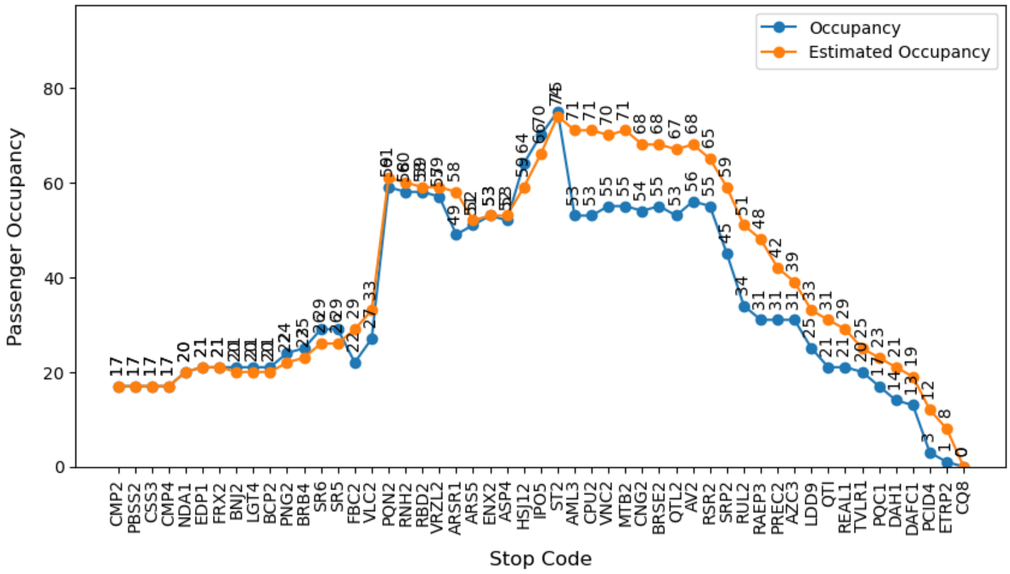
The variance explanatory power is considerably lower for exits (See Table 5.2). This happens because the amount of passenger exits by stop is often zero or small values. Since the EVS considers the variance between the predicted and actual values, in the presence of many zeros or small values it is expected that the model's ability to explain the variance will be limited, resulting in low EVS scores.

	Exits				
	MAE	**MAN**	**RMSE**	**RMSN**	**EVS**
AVR	0,920	0,117	1,573	0,196	0,270
STD	0,377	0,033	0,644	0,041	0,199
MIN	0,333	0,0525	0,626	0,101	-0,439
MAX	2,471	0,233	4,082	32,489	0,733

Table 5.2: Overall Exits Estimation Results

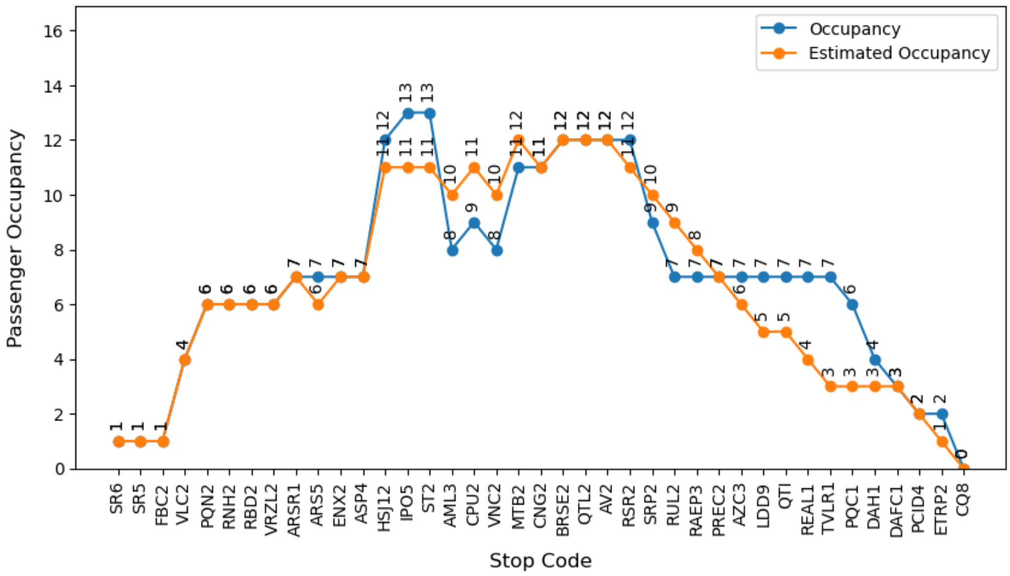
The trip with the most passengers, 147 in this case, occurred in the Last Val period of a working day, during Summer season. It got an occupancy ESV score of 0,92 and its results can be analyzed in Figure 5.4. The occupation peak was correctly estimated in the correct stop.

Figure 5.4: Trip with the most passengers: Actual and Predicted Occupancy



The trip with the least passengers out of the evaluated, which were 20 as stipulated, occurred in the Last Val period of a weekend/Holiday day, during August season. The ESV score for this trip was 0,88, and its results can be analyzed in Figure 5.5.

Figure 5.5: Trip with the least passengers: Actual and Predicted Occupancy



5.2.1 Outward vs Return Trips

In this section, the results of the occupancy estimation model are analyzed based on the distinction between outward and return trips. The aim is to investigate whether any notable differences in the model's performance can be observed for these two types of trips.

Table 5.3 presents the estimation results for outward (Direction 1) and return (Direction 2) trips separately. The table includes the number of trips in each direction, as well as the average values of the metrics used to evaluate the model's performance.

	Occupation			**Exits**		
Direction (Nr)	**MAN**	**RMSN**	**EVS**	**MAN**	**RMSN**	**EVS**
1 (152)	0,105	0,144	0,815	0,123	0,203	0,218
2 (140)	0,121	0,167	0,789	0,110	0,187	0,325

Table 5.3: Estimation Results by Direction

For outward trips (Direction 1), the model achieves an occupancy NMAE of approximately 0,105, and an occupancy NRMSE of 0,14,4. The EVS value suggests that the model can account for approximately 81.5% of the variability in the actual occupancy for outward trips.

For return trips (Direction 2), the model shows slightly higher estimation errors. The occupancy NMAE is approximately 0,121, and the NRMSE is around 0,167. The EVS value indicates that the model can explain approximately 78.9% of the variability in the actual occupancy for return trips.

The model was expected to perform better for direction 2, as return trips have less stops, increasing the odds of correctly estimating the exits. Additionally, direction 2 has a single stop with no direction 1 correspondence, meaning no real information on opposite direction entries, compared to 7 stops in the inverse condition. However, all the metrics indicate that the model's accuracy is higher for outward trips (Direction 1) compared to return trips (Direction 2). The causes for this difference are not clear, and may be due to variations in passenger patterns, demand, or other factors.

5.2.2 Line Variant

In this section, the results of the occupancy estimation model are examined based on different line variants. The purpose is to determine if there are any notable variations in the model's performance across these variants.

Table 5.4 provides a summary of the estimation results for each line variant. The number of trips for each variant is specified, along with the corresponding metrics. The variants are sorted from trips with more stops to trips with less stops, and for all of the different variant types (Complete, Var1 and Var2), the first line corresponds to the direction 1 trips and the second to direction 2.

		Occupation			**Exits**		
Type	**Variant (Nr)**	**MAN**	**RMSN**	**EVS**	**MAN**	**RMSN**	**EVS**
Complete	CMP-CQ (67)	0,109	0,148	0,792	0,111	0,189	0,226
	CQ-CMP (63)	0,158	0,214	0,720	0,107	0,185	0,269
Var1	SR-CQ (51)	0,100	0,137	0,859	0,133	0,215	0,155
	CQ-SR (44)	0,090	0,127	0,850	0,110	0,189	0,325
Var2	SR-LDD (34)	0,105	0,146	0,795	0,131	0,211	0,298
	LDD-SR (33)	0,092	0,129	0,841	0,115	0,190	0,433

Table 5.4: Estimation Results by Variant

The occupancy errors remain similar regardless of the variant, with the exception of the CQ-CMP trips, from direction 2 that had a occupancy NMAE of 0,158 and a occupancy NRMSE of 0,214. Otherwise, the occupancy NMAE values range from 0,090 to 0,109, while the occupancy NRMSE ranges from 0,127 to 0,148.

The model seems to capture the underlying patterns and variations from Variant 1 trips very well, as indicated by an EVS value of 0,850 for SR-CQ and 0,859 for CQ-SR. The variants CMP-CQ and SR-LDD have EVSs of 0,792 and 0,795 respectively, having performances within the average of the model (overall EVS of 0,803), while LDD-SR is slightly better fitted than the average, with an EVS of 0,841. The complete direction 2 trips (CQ-CMP) have an uncharacteristically low EVS, comparing to all other trips, of 0,718.

Looking at the results by variant, it is noticeable that the model only performs better for direction 1 for complete trips. For the rest of the variants (Var1 and Var2), direction 2 shows better results. Once again, CQ-CMP stands out from the rest of the variants, with results so atypical that the analysis solely based on direction led to the conclusion that outward trips performed better (Direction 1).

The rest of the direction 1 variants (CQ-SR and LDD-SR) have favorable results. One explanation for the CQ-CMP outcomes could be that the model performs particularly badly in the 9 stations exclusive to CQ-CMP (from SR to CMP).

5.2.3 Working Day vs Weekend Trips

In this section, the results of the occupancy estimation model are analyzed based on the differentiation between working day and weekend trips. The aim is to assess whether there are any notable discrepancies in the model's performance for these two categories and to investigate the impact of incorporating this factor into the model.

Table 5.5 provides a summary of the estimation results for both working day and weekend trips. The table includes the number of trips for each category, along with the corresponding metrics. For trips within the same Weekday category, the first line showcases the outcomes of the developed model, while the second line displays the metrics when the day of the week is not considered by the model.

		Occupation			**Exits**		
Weekday (Nr)	**Matched to**	**MAN**	**RMSN**	**EVS**	**MAN**	**RMSN**	**EVS**
Work Day (256)	Work Day	0,112	0,154	0,802	0,117	0,197	0,275
	All	0,112	0,154	0,805	0,117	0,197	0,275
Weekend (36)	Weekend	0,115	0,159	0,808	0,110	0,188	0,234
	All	0,119	0,165	0,795	0,111	0,189	0,221

Table 5.5: Estimation Results by Weekday

While working day trips have an occupancy NMAE of approximately 0,112, this value is 0,119 for weekend. The estimation results for working days indicate an occupation NRMSE of 0,154 while weekend trips indicate 0,159. As for occupancy EVS values, they are very similar, with 0,808 for weekends and 0,802 for working days. Overall, the model seems to perform very similarly regardless of the weekday.

The weekday of a trip was included as a factor in the occupancy estimation model, as the Data analysis revealed differentiating passenger travel patterns. The results presented in Table 5.5 show the impact of this factor in the model.

By comparison, when the day of the week is not taken into account by the model (labeled as "Matched to" "All days"), the occupation errors are not affected. Similarly, the occupation EVS remains very similar, dropping slightly for weekends, and increasing slightly for working days.

In conclusion, the results suggest that the weekday factor does not affect the accuracy and explanatory power of the model considerably.

5.2.4 Daily Period

In this section, the results of the occupancy estimation model are examined based on different daily periods. The objective is to assess whether there are any notable variations in the model's performance across these periods and to explore the effects of incorporating this factor into the model.

Table 5.6 provides a summary of the estimation results for each daily period. As previously defined, the various periods refer to daily time frames. *First Peak*: between 7:00 and 8:00; *Last Peak*: between 17:00 and 18:00; *First Val*: between 5:00 and 11:00; *Last Val*: between 15:00 and 00:00; *Undefined*: between 11:00 and 15:00; *Total*: all day. The table includes information regarding the number of trips for each period, as well as the corresponding metrics. For trips within the same period category, the first line represents the outcomes of the developed model, while the second line displays the metrics when the daily period is not considered.

		Occupation			**Exits**		
Period (Nr)	**Matched to**	**MAN**	**RMSN**	**EVS**	**MAN**	**RMSN**	**EVS**
Undefined (71)	Total	0,117	0,156	0,810	0,117	0,193	0,314
	Total	0,117	0,156	0,810	0,117	0,193	0,314
First Val (65)	Last Val	0,112	0,156	0,784	0,108	0,189	0,275
	Total	0,121	0,171	0,747	0,113	0,197	0,220
Last Val (91)	First Val	0,101	0,140	0,838	0,121	0,194	0,308
	Total	0,103	0,140	0,844	0,114	0,186	0,355
First Peak (23)	Last Peak	0,141	0,200	0,704	0,107	0,203	0,109
	Total	0,149	0,214	0,682	0,109	0,207	0,080
Last Peak (42)	First Peak	0,116	0,157	0,797	0,126	0,208	0,191
	Total	0,122	0,163	0,801	0,119	0,204	0,213

Table 5.6: Estimation Results by Period

The results are overall similar, with slight variations in the occupancy estimation performance across different daily periods. The Fist Peak period stands out, with NMAE and NRMSE occupancy errors of 0,141 and 0,200 respectively, and a EVS that indicates less occupancy variability explained by the model, at a value of approximately 71,9%. The Last Val occupancy EVS is slightly higher than the overall average, at a value of 0,838. The performances of the remaining periods resemble those of the overall model.

When the daily period is not taken into account by the model (labeled as "Matched to" "Total"), the occupation errors increase consistently for all periods, while the EVS occupancy's scores overall decrease. The Last Val and Last Peak periods are an exception to this pattern. For the Last Val period the exclusion of the daily period factor from the model leads to a slight EVS increase of 0,006, while for the Last Peak period this increase is of 0,004.

Overall, the findings suggest that with the exception of the First Peak period, the model behaves similarly regardless of trip period. Additionally, incorporating the daily period factor in the occupancy estimation model appears relevant. By considering the specific characteristics and dynamics of different periods throughout the day, the model can better capture and predict passenger behavior, resulting in improved accuracy and performance in estimating occupancy.

5.2.5 Seasonality

In order to evaluate the performance of the model trips from different seasons, and assess the significance of the inclusion of this factor on the model's performance, the model was run, both including and excluding this element. Since the Anda generated trips were validations grouped by a number of factors, including the month, and the school Holidays season only started on December 19^{th}, this season was not evaluated. For the case of Summer, since it started on June 15^{th}, only Anda trips from the month of July were matched to this season.

Table 5.7 presents a summary of the estimation results for each specific season of the year. The table provides details regarding the number of trips associated to each season, and their corresponding metrics. For trips within the same season, the first line, Matched to the season itself, corresponds to the developed model, while the second, Matched to the Entire Year, corresponds to the exclusion of the Season factor.

		Occupation			**Exits**		
Season (Nr)	**Matched to**	**MAN**	**RMSN**	**EVS**	**MAN**	**RMSN**	**EVS**
Regular (249)	Regular	0,115	0,158	0,797	0,118	0,198	0,268
	Total	0,115	0,157	0,798	0,118	0,198	0,269
Summer (21)	Summer	0,102	0,141	0,805	0,110	0,186	0,270
	Total	0,100	0,140	0,807	0,109	0,185	0,274
August (18)	August	0,089	0,126	0,872	0,103	0,178	0,294
	Total	0,093	0,132	0,858	0,105	0,182	0,260

Table 5.7: Estimation Results by Season

The performance of the model for trips during Regular season and during the Summer are pretty similar, and follow the overall model results. The August trips however, perform very well, with the lowest errors of the entire analysis (occupation NMAE of 0,089 and NRMSE of 0,126) and the highest ESV, indicating that the model can explain 87.2% of the variability in the actual occupancy.

When examining the impact of including the seasonality factor on the estimation results, it is observed that the model performance remains consistently pretty much the same. The results indicate that considering seasonality as a factor in the occupancy estimation model does not impact considerably the results, even for the month of August, meaning that the better performance could be a coincidence.

5.2.6 Discussion

The model was expected to perform better when taking into account daily period, weekday, and season. Not only these factors demonstrated significant variations in passenger behaviour in the data analysis process, they are also very commonly associated to seasonality, capturing unique characteristics. However the exclusion of these specifications from the model affected it very little.

Overall, the results show consistently a good performance of the model. The occupancy errors are considerably low averaging on a maximum error of 11,3% of the maximum trip occupancy, at any given station. This value increases to 15,5% in the case of the RMSE, that focuses on the magnitude of the errors, penalizes larger errors more heavily. The model's ESV indicates that the estimated occupancy accounts for approximately 80.3% of the variability in the actual occupancy.

5.3 Limitations

This section discusses the limitations associated with the occupancy estimation model developed in this study. The model's performance and validation method are evaluated, acknowledging potential constraints that may impact its accuracy and reliability.

5.3.1 Model Limitations

It is essential to acknowledge the limitations of the occupancy estimation model:

Errors in data: The accuracy of the model heavily relies on the quality and reliability of the data used, including the entry and exit information. While the data was extensively cleaned, it still includes inaccuracies that could affect the model's performance and the validity of its predictions. For instance regarding the trips variant identification.

Specificity of the case study: The model was developed based on the characteristics and patterns observed in Line 205, which has an overall urban profile and function. It is important to recognize that the same model may not yield similar results for bus lines that are less commonly used for commuting or have different operating characteristics.

Model Assumptions and Sensitivity:

The model relies on several assumptions derived from the analysis of the entry data. While these assumptions are based on observed patterns, it is possible that they may not fully capture all variations in passenger behavior. Consequently, the model's estimates may be subject to deviations from the actual values in certain scenarios. Additionally, the accuracy of the occupancy estimates is highly dependent on the validity of the assumptions made. If there are significant deviations or variations in passenger behavior that are not adequately captured by the model's assumptions, the reliability of the model's predictions may be compromised.

External factors not taken into account:

The model assumes that passenger occupancy follows consistent patterns, and it does not take into account external factors such as weather conditions, local events, or other dynamic influences that could impact passenger behavior and occupancy levels.

Overall, the bus line occupancy estimation model developed using only entry data represents a significant advancement in accurately estimating the number of passengers on board. By leveraging patterns and characteristics of boarding behavior, the model provides valuable information for optimizing bus line operations and resource allocation. However, the limitations outlined above need to be considered when interpreting and applying the model's results.

5.3.2 Validation Limitations

It is crucial to recognize the limitations of the validation method employed in this study. The validation process was conducted based on the available data and the assumptions made during the model's development.

The accuracy of the ground truth values used for validation introduces a level of uncertainty in the evaluation. It is important to reiterate that the trips evaluated in the validation process do not correspond to actual bus trip journey information, but rather simulated trips generated using the Anda dataset, which contains actual boarding and alighting information. While the simulated trips aim to capture similar travel patterns, there may still be variations and discrepancies compared to real-world bus trips, which can impact the accuracy of the model's validation process.

The validation of the study evaluated a total of 292 trips, with certain categories having as few as 18 evaluated trips, which is relatively low. This limited sample size could introduce statistical uncertainties and affect the overall robustness of the validation results.

Despite these limitations, the validation method provided a comprehensive assessment of the bus line occupancy estimation model. The obtained results served as a basis for evaluating the model's performance, identifying potential areas for improvement, and ensuring its suitability for estimating occupancy based solely on entry data.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In conclusion, the various aspects of modeling occupancy estimation in public transportation systems have been explored throughout this thesis. Through an extensive review of the literature and analysis of different methodologies, valuable insights have been gained into the challenges and opportunities associated with accurately estimating passenger loads.

A public transport occupancy estimation model was successfully developed, following the CRISP-DM methodology. The first step, business understanding, consisted in the exploration of the case study context, in addition to the state-of-art establishment, leading to valuable insights that informed the project. Followed by the data understanding, which involved the description, familiarisation with the datasets, and a lot of analysis in order to extract knowledge. The data cleaning process was very meticulous and resulted in a very high rate of errors mitigation. Using all of the information accumulated about the data in this process, the modelling phase was conducted, followed by the Validation phase. All of the research objectives were successfully fulfilled, and the conducted work was rigorously described in this document.

As expected, this process was not linear, new issues in the data were constantly discovered, leading to a re-visitation of data cleaning. Similarly, the conduction of the validation informed the modeling process, that was adjusted. The development of the project was very iterative.

This thesis revealed significant insights regarding travel patterns in the case study, particularly regarding the correlation between opposite directions' demands for both passenger boardings and passenger alighting. And once again, resulted in a robust AFC data based occupancy estimation model, that successfully captures the complex dynamics of passenger flow and accurately estimates vehicle occupancy levels. The validation of the model demonstrated its good performance and reliability, indicating its potential as a valuable tool for transport planning and decision-making processes.

Overall, the main findings of this research contribute to the body of knowledge of ITS, and in particular passenger flow estimation using AFC systems' data. The research outcomes provide insights into passenger behavior, travel patterns, and the accurate estimation of vehicle occupancy.

6.2 Future Work

This section discusses the future work and potential improvements that can be explored in the context of estimating occupancy in bus trips. By addressing these aspects, it is anticipated that the accuracy and performance of the occupancy estimation model can be further improved, leading to more reliable predictions and valuable insights for optimizing bus operations.

Firstly, the current model does not take into account the type of ticket (monthly subscriptions, temporary subscriptions, or occasional ticket) as a factor. Significant differences between the different types of passengers were not revealed during the data analysis process; however, incorporating this information could potentially improve the accuracy of the model.

Additionally, the Anda dataset, which was found to be representative of the Andante dataset, was exclusively used for model validation. However, as it contains unique information regarding passengers exits, it can also serve the following purposes:

Feature Engineering: The analysis of the Anda dataset can aid in identifying relevant features or variables that correlate with occupancy. By exploring factors such as trip duration, time of day, day of the week, specific stops, or other contextual information captured in Anda, patterns and trends related to occupancy can be uncovered. These insights can inform the feature engineering process for developing more accurate occupancy estimation models.

Model Calibration: The Anda dataset can be leveraged for calibrating and fine-tuning occupancy estimation models. By incorporating the Anda data into the modeling process, adjustments can be made to the model's parameters or algorithms to align the estimated occupancy values more closely with the ground truth data. This calibration step helps enhance the accuracy and reliability of the models, improving their performance in estimating occupancy for the entire bus line.

Furthermore, exploring advanced machine learning techniques may further enhance the predictive capabilities of the model.

References

[1] “A multi-pattern deep fusion model for short-term bus passenger flow forecasting”. In: *Applied Soft Computing* 58 (2017), pp. 669–680. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2017.05.011.

[2] Azalden Alsger et al. “Public transport trip purpose inference using smart card fare data”. In: *Transportation Research Part C: Emerging Technologies* 87 (2018), pp. 123–137. ISSN: 0968-090X. DOI: 10.1016/j.trc.2017.12.016.

[3] Azalden Alsger et al. “Validating and improving public transport origin–destination estimation algorithm using smart card fare data”. In: *Transportation Research Part C: Emerging Technologies* 68 (2016), pp. 490–506. ISSN: 0968-090X. DOI: 10.1016/j.trc.2016.05.004.

[4] Azalden A Alsger et al. “Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix”. In: *Transportation Research Record* 2535.1 (2015), pp. 88–96. DOI: 10.3141/2535-10.

[5] Fundo Ambiental. *Programa de Apoio à Redução do Tarifário dos Transportes Públicos (PART)*. 2020. URL: https://www.fundoambiental.pt/avisos-anteriores/apoios-20211/mitigacao-das-alteracoes-climaticas1/programa-de-apoio-a-reducao-do-tarifario-dos-transportes-publicos-part.aspx (visited on 06/08/2023).

[6] I J Amin et al. “Automated people-counting by using low-resolution infrared and visual cameras”. In: *Measurement* 41.6 (2008), pp. 589–599. ISSN: 0263-2241. DOI: 10.1016/j.measurement.2007.02.010.

[7] AMPorto. *Alargamento Andante*. 2019. URL: http://portal.amp.pt/pt/noticia/481 (visited on 06/08/2023).

[8] AMPorto. *Caracterização da AMP*. 2020. URL: http://portal.amp.pt/pt/4/stats/territorio/#FOCO_4 (visited on 03/22/2023).

[9] AMPorto. *N12 Qualificação Urbana da Circunvalação*. 2017. URL: http://portal.amp.pt/pt/3/projectos20/223 (visited on 06/08/2023).

[10] AMPorto. *Operadores AMP*. 2020. URL: http://autoridade.amp.pt/pt-pt/home/transporte-na-amp/transportes/ (visited on 03/22/2023).

[11] Andante. *Transportes Intermodais do Porto*. URL: https://www.linhAndante.com/web/tip/tip-transportes-intermodais-do-porto (visited on 03/22/2023).

[12] Behrang Assemi et al. "Improving alighting stop inference accuracy in the trip chaining method using neural networks". In: *Public Transport* 12 (Mar. 2020), pp. 89–121. DOI: 10.1007/s12469-019-00218-9.

[13] James Barry et al. "Origin and Destination Estimation in New York City with Automated Fare System Data". In: *Transportation Research Record* 1817 (Jan. 2002), pp. 183–187. DOI: 10.3141/1817-24.

[14] James J. Barry, Robert Freimer, and Howard Slavin. "Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City". In: *Transportation Research Record* 2112.1 (2009), pp. 53–61. DOI: 10.3141/2112-07.

[15] M Belloc et al. "Detection of people boarding/alighting a metropolitan train using computer vision". In: (2018). DOI: 10.1049/cp.2018.1281.

[16] Nicola Bernini et al. "An embedded system for counting passengers in public transportation vehicles". In: *2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA)*. 2014, pp. 1–6. DOI: 10.1109/MESA.2014.6935562.

[17] Maurizio Bielli, Alessandro Bielli, and Riccardo Rossi. "Trends in Models and Algorithms for Fleet Management". In: *Procedia - Social and Behavioral Sciences* 20 (2011), pp. 4–18. ISSN: 1877-0428. DOI: 10.1016/j.sbspro.2011.08.004.

[18] Anne-Sarah Briand et al. "Analyzing year-to-year changes in public transport passenger behaviour using smart card data". In: *Transportation Research Part C: Emerging Technologies* 79 (2017), pp. 274–289. ISSN: 0968-090X. DOI: 10.1016/j.trc.2017.03.021.

[19] Chao-Ho Chen et al. "People Counting System for Getting In/Out of a Bus Based on Video Processing". In: *2008 Eighth International Conference on Intelligent Systems Design and Applications*. Vol. 3. 2008, pp. 565–569. DOI: 10.1109/ISDA.2008.335.

[20] Juan Chen et al. "Automatic head detection for passenger flow analysis in bus surveillance videos". In: *2012 5th International Congress on Image and Signal Processing*. 2012, pp. 143–147. DOI: 10.1109/CISP.2012.6469669.

[21] Yingjie Chen, Liquan Zhang, and Jia Wang. "Passenger detection for subway transportation based on video". In: *2014 10th International Conference on Natural Computation (ICNC)*. 2014, pp. 720–724. DOI: 10.1109/ICNC.2014.6975925.

[22] J.P. Deparis et al. "A device for counting passengers making use of two active linear cameras: comparison of algorithms". In: *1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems (Cat. No.96CH35929)*. Vol. 3. 1996, 1629–1634 vol.3. DOI: 10.1109/ICSMC.1996.565339.

[23] Andrei Donțu et al. “Weigh-in-motion sensors and traffic monitoring systems - Sate of the art and development trends”. In: *IOP Conference Series: Materials Science and Engineering* 997 (Dec. 2020), p. 012113. DOI: 10.1088/1757-899X/997/1/012113.

[24] Instituto Nacional de Estatistica. *Resident population*. 2021. URL: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&contecto=pi&indOcorrCod=0011628&selTab=tab0 (visited on 03/22/2023).

[25] Hamed Faroqi, Mahmoud Mesbah, and Jiwon Kim. “Applications of transit smart cards beyond a fare collection tool: A literature review”. In: *Advances in Transportation Studies* 45 (July 2018), pp. 107–122. DOI: 10.4399/978255166098.

[26] Janine M Farzin. “Constructing an Automated Bus Origin–Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil”. In: *Transportation Research Record* 2072.1 (2008), pp. 30–37. DOI: 10.3141/2072-04.

[27] Horst E Gerland and Kurt Sutter. “Automatic passenger counting (apc): Infra-red motion analyzer for accurate counts in stations and rail, light-rail and bus operations”. In: *1999 APTA Bus Conference, ProceedingsAmerican Public Transportation Association*. 1999. URL: https://trid.trb.org/view/504735.

[28] Jason Gordon et al. “Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2343 (Dec. 2013), pp. 17–24. DOI: 10.3141/2343-03.

[29] Luke Haywood and Martin Koning. “Estimating crowding costs in public transport”. In: (2013). DOI: 10.2139/ssrn.2256332.

[30] Léonie Heydenrijk-Ottens et al. “Supervised learning: Predicting passenger load in public transport”. In: 2018.

[31] Joana Hora et al. “Estimation of Origin-Destination matrices under Automatic Fare Collection: the case study of Porto transportation system”. In: *Transportation Research Procedia* 27 (Jan. 2017), pp. 664–671. DOI: 10.1016/j.trpro.2017.12.103.

[32] Rong Hu, Yi-Chang Chiu, and Chih-Wei Hsieh. “Crowding Prediction on Mass Rapid Transit systems using a Weighted Bidirectional Recurrent Neural Network”. In: *IET Intelligent Transport Systems* 14 (Mar. 2020). DOI: 10.1049/iet-its.2018.5542.

[33] Saddam Hussain Khan et al. “Passenger detection and counting during getting on and off from Public transport systems”. In: (2020). DOI: 10.35453/NEDJR-ASCN-2019-0016.

[34] Erik Jenelius. “Data-Driven Metro Train Crowding Prediction Based on Real-Time Load Data”. In: *IEEE Transactions on Intelligent Transportation Systems* 21.6 (2020), pp. 2254–2265. DOI: 10.1109/TITS.2019.2914729.

[35] Feifan Jia et al. "A Deep Learning-Based Hybrid Model for Short-term Subway Passenger Flow Prediction using AFC Data". In: *IET Intelligent Transport Systems* 13 (Nov. 2019). DOI: 10.1049/iet-its.2019.0112.

[36] Xianyu Jianchuan. "Research on the interdependencies between commute mode choice and trip chaining behavior". In: *2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings*. 2013, pp. 2042–2046. DOI: 10.1109/ICMSE.2013.6586545.

[37] Alireza Khani. *Models and Solution Algorithms for Transit and Intermodal Passenger Assignment (Development of FAST-TrIPs Model)*. 2013. URL: http://hdl.handle.net/10150/306074.

[38] Louahdi Khoudour, Luc Duvieubourg, and Jean-Pierre Deparis. "Real-time passenger counting by active linear cameras". In: *Real-Time Imaging*. Ed. by Phillip A Laplante, Alexander D Stoyenko, and Divyendu Sinha. Vol. 2661. International Society for Optics and Photonics. SPIE, 1996, pp. 106–117. DOI: 10.1117/12.234643.

[39] Jiwon Kim, Jonathan Corcoran, and Marty Papamanolis. "Route choice stickiness of public transport passengers: Measuring habitual bus ridership behaviour using smart card data". In: *Transportation Research Part C: Emerging Technologies* 83 (2017), pp. 146–164. ISSN: 0968-090X. DOI: 10.1016/j.trc.2017.08.005.

[40] Denis Klauser, Günter Bärwolff, and Hartmut Schwandt. "A TOF-based automatic passenger counting approach in public transportation systems". In: *AIP Conference Proceedings* 1648.1 (2015), p. 850113. DOI: 10.1063/1.4913168.

[41] Pavol Kuchár et al. "Noninvasive Passenger Detection Comparison Using Thermal Imager and IP Cameras". In: *Sustainability* 13.22 (2021). ISSN: 2071-1050. DOI: 10.3390/su132212928.

[42] Pavol Kuchár et al. "Passenger Occupancy Estimation in Vehicles: A Review of Current Methods and Research Challenges". In: *Sustainability* 15 (Jan. 2023), p. 1332. DOI: 10.3390/su15021332.

[43] Pramesh Kumar, Alireza Khani, and Qing He. "A robust method for estimating transit passenger trajectories using automated data". In: *Transportation Research Part C: Emerging Technologies* 95 (Oct. 2018), pp. 731–747. DOI: 10.1016/j.trc.2018.08.006.

[44] Takahiko Kusakabe and Yasuo Asakura. "Behavioural data mining of transit smart card data: A data fusion approach". In: *Transportation Research Part C: Emerging Technologies* 46 (2014), pp. 179–191. ISSN: 0968-090X. DOI: 10.1016/j.trc.2014.05.012.

[45] Antonio Lazaro et al. "Seat-Occupancy Detection System and Breathing Rate Monitoring Based on a Low-Cost mm-Wave Radar at 60 GHz". In: *IEEE Access* 9 (2021), pp. 115403–115414. DOI: 10.1109/ACCESS.2021.3105390.

[46] Inmook Lee et al. "Travel pattern-based bus trip origin-destination estimation using smart card data". en. In: *PLoS One* 17.6 (July 2022). DOI: 10.1371/journal.pone.0270346.

[47] Sang Gu Lee and Mark Hickman. “Trip purpose inference using automated fare collection data”. In: *Public Transport* 6 (2014), pp. 1–20. DOI: 10.1007/s12469-013-0077-5.

[48] P Lengvenis et al. “Application Of Computer Vision Systems For Passenger Counting In Public Transport”. In: *Elektronika ir Elektrotechnika* 19.3 (2013), pp. 69–72. DOI: 10.5755/j01.eee.19.3.1232.

[49] Feng Li et al. “Automatic Passenger Counting System for Bus Based on RGB-D Video”. In: *Proceedings of the 2nd Annual International Conference on Electronics, Electrical Engineering and Information Science (EEEIS 2016)*. Atlantis Press, Dec. 2016, pp. 209–220. ISBN: 978-94-6252-320-3. DOI: 10.2991/eeeis-16.2017.29.

[50] Daniele Liciotti et al. “An Intelligent RGB-D Video System for Bus Passenger Counting”. In: *Intelligent Autonomous Systems 14*. Ed. by Weidong Chen et al. Vol. 531. Cham: Springer International Publishing, Feb. 2017, pp. 473–484. ISBN: 978-3-319-48036-7. DOI: 10.1007/978-3-319-48036-7_34.

[51] Jenny Liu et al. “Human Occupancy Detection via Passive Cognitive Radio”. In: *Sensors* 20.15 (2020). ISSN: 1424-8220. DOI: 10.3390/s20154248.

[52] Wiktoria Loga-Ksieska, Krzysztof Brzozowski, and Artur Ryguła. “A Method for Estimating the Occupancy Rates of Public Transport Vehicles Using Data from Weigh-In-Motion Systems”. In: *Challenge of Transport Telematics: 16th International Conference on Transport Systems Telematics, TST 2016, Katowice-Ustroń, Poland, March 16–19, 2016, Selected Papers*. Vol. 640. Nov. 2016, pp. 426–435. ISBN: 978-3-319-49645-0. DOI: 10.1007/978-3-319-49646-7_36.

[53] Jonathan Samuel Lumentut, Fergyanto E Gunawan, and Diana. “Evaluation of Recursive Background Subtraction Algorithms for Real-Time Passenger Counting at Bus Rapid Transit System”. In: *Procedia Computer Science* 59 (2015), pp. 445–453. ISSN: 1877-0509. DOI: 10.1016/j.procs.2015.07.565.

[54] Xiaolei Ma et al. “Mining smart card data for transit riders’ travel patterns”. In: *Transportation Research Part C: Emerging Technologies* 36 (2013), pp. 1–12. ISSN: 0968-090X. DOI: 10.1016/j.trc.2013.07.010.

[55] Ubaid Mehmood et al. “Occupancy Estimation using WiFi: A Case Study for Counting Passengers on Busses”. In: *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*. 2019, pp. 165–170. DOI: 10.1109/WF-IoT.2019.8767350.

[56] Lars Mikkelsen et al. “Public transport occupancy estimation using WLAN probing”. In: *2016 8th International Workshop on Resilient Networks Design and Modeling (RNDM)*. 2016, pp. 302–308. DOI: 10.1109/RNDM.2016.7608302.

[57] Satarupa Mukherjee et al. “Anovel framework for automatic passenger counting”. In: *2011 18th IEEE International Conference on Image Processing*. 2011, pp. 2969–2972. DOI: 10.1109/ICIP.2011.6116284.

[58] Marcela Munizaga and Carolina Palma. "Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile". In: *Transportation Research Part C-Emerging Technologies* 24 (Oct. 2012), pp. 9–18. DOI: 10.1016/j.trc.2012.01.007.

[59] Marcela Munizaga et al. "Validating travel behavior estimated from smartcard data". In: *Transportation Research Part C: Emerging Technologies* 44 (July 2014), pp. 70–79. DOI: 10.1016/j.trc.2014.03.008.

[60] Hayato Nakashima, Ismail Arai, and Kazutoshi Fujikawa. "Passenger Counter Based on Random Forest Regressor Using Drive Recorder and Sensors in Buses". In: *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2019, pp. 561–566. DOI: 10.1109/PERCOMW.2019.8730761.

[61] Hayato Nakashima, Ismail Arai, and Kazutoshi Fujikawa. "Proposal of a Method for Estimating the Number of Passengers with Using Drive Recorder and Sensors Equipped in Buses". In: *2018 IEEE International Conference on Big Data (Big Data)*. 2018, pp. 5396–5398. DOI: 10.1109/BigData.2018.8621983.

[62] A.S.A. Nasir, N.K.A. Gharib, and H. Jaafar. "Automatic Passenger Counting System Using Image Processing Based on Skin Colour Detection Approach". In: *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*. 2018, pp. 1–8. DOI: 10.1109/ICASSDA.2018.8477628.

[63] Neema Nassir et al. "Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System". In: *Transportation Research Record: Journal of the Transportation Research Board* 2263 (Dec. 2011), pp. 140–150. DOI: 10.3141/2263-16.

[64] António A. Nunes, Teresa Galvão Dias, and João Falcão e Cunha. "Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation". In: *IEEE Transactions on Intelligent Transportation Systems* 17.1 (2016), pp. 133–142. DOI: 10.1109/TITS.2015.2464335.

[65] Tim van Oosterhout, Sander Bakkes, and Ben J A Kröse. "Head Detection in Stereo Data for People Counting and Segmentation". In: *VISAPP 2011 - Proceedings of the Sixth International Conference on Computer Vision Theory and Applications, Vilamoura, Algarve, Portugal, 5-7 March, 2011*. Ed. by Leonid M Mestetskiy and José Braz. SciTePress, 2011, pp. 620–625. URL: https://www.researchgate.net/publication/221415646_Head_Detection_in_Stereo_Data_for_People_Counting_and_Segmentation.

[66] Kevin Pasini et al. "LSTM encoder-predictor for short-term train load forecasting". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III*. Springer. 2020, pp. 535–551. DOI: 10.1007/978-3-030-46133-1_32.

[67] Jau-Woei Perng et al. "The design and implementation of a vision-based people counting system in buses". In: *2016 International Conference on System Science and Engineering (ICSSE)*. 2016, pp. 1–3. DOI: 10.1109/ICSSE.2016.7551620.

[68] Federação Académica do Porto. *Autocarros-vaivém gratuitos: Queima das Fitas do Porto 2019*. 2019. URL: https://www.fap.pt/noticias/autocarros-vaivem-gratuitos-queima-das-fitas-do-porto-2019 (visited on 06/13/2023).

[69] Metro do Porto. *Transportes Intermodais do Porto*. 2018. URL: https://www.metrodoporto.pt/frontoffice/pages/379?news_id=311 (visited on 06/08/2023).

[70] Xiao Qian, Xiao Yu, and Chen Fa. "The passenger flow counting research of subway video based on image processing". In: *2017 29th Chinese Control And Decision Conference (CCDC)*. 2017, pp. 5195–5198. DOI: 10.1109/CCDC.2017.7979418.

[71] Alam Rahmatulloh et al. "Applied Internet of Things (IoT): The Prototype Bus Passenger Monitoring System Using PIR Sensor". In: *2020 8th International Conference on Information and Communication Technology (ICoICT)*. 2020, pp. 1–6. DOI: 10.1109/ICoICT49345.2020.9166420.

[72] Steve Robinson et al. "Methods for pre-processing smartcard data to improve data quality". In: *Transportation Research Part C: Emerging Technologies* 49 (Dec. 2014). DOI: 10.1016/j.trc.2014.10.006.

[73] Claudio Roncoli, Ektoras Chandakas, and Ioannis Kaparias. "Estimating on-board passenger comfort in public transport vehicles using incomplete automatic passenger counting data". In: *Transportation Research Part C: Emerging Technologies* 146 (Jan. 2023), p. 103963. DOI: 10.1016/j.trc.2022.103963.

[74] Michael Siebert and David Ellenberger. "Validation of automatic passenger counting: introducing the t-test-induced equivalence test". In: *Transportation* (Dec. 2020). DOI: 10.1007/s11116-019-09991-9.

[75] "Statistical analysis of passenger-crowding in bus transport network of Harbin". In: 490 (2018), pp. 426–438. DOI: 10.1016/j.physa.2017.08.004.

[76] Hermann Sterner, Wolfgang Aichholzer, and Matthias Haselberger. "Development of an Antenna Sensor for Occupant Detection in Passenger Transportation". In: *Procedia Engineering* 47 (2012), pp. 178–183. ISSN: 1877-7058. DOI: 10.1016/j.proeng.2012.09.113.

[77] Alejandro Tirachini et al. "Estimation of crowding discomfort in public transport: Results from Santiago de Chile". In: *Transportation Research Part A: Policy and Practice* 103 (June 2017). DOI: 10.1016/j.tra.2017.06.008.

[78] Sociedade de Transportes Colectivos do Porto. *Linhas*. 2023. URL: https://www.stcp.pt/pt/viajar/linhas/ (visited on 03/22/2023).

[79] Martin Trépanier, Nicolas Tranchant, and Robert Chapleau. "Individual trip destination estimation in a transit smart card automated fare collection system". In: *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 11.1 (2007), pp. 1–14. DOI: 10.1080/15472450601122256.

[80] Túlio Vieira et al. "Public Transport Occupancy Estimation using WLAN Probing and Mathematical Modeling". In: *Transportation Research Procedia* 48 (Jan. 2020), pp. 3299–3309. DOI: 10.1016/j.trpro.2020.08.122.

[81] Wei Wang, John Attanucci, and Nigel Wilson. "Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems". In: *Journal of Public Transportation* 14 (Dec. 2011), pp. 131–150. DOI: 10.5038/2375-0901.14.4.7.

[82] Libo Wu and Ya Wang. "A Low-Power Electric-Mechanical Driving Approach for True Occupancy Detection Using a Shuttered Passive Infrared Sensor". In: *IEEE Sensors Journal* 19.1 (2019), pp. 47–57. DOI: 10.1109/JSEN.2018.2875659.

[83] Tarek Yahiaoui, Louahdi Khoudour, and Cyril Meurie. "Real-time passenger counting in buses using dense stereovision". In: *Journal of Electronic Imaging* 19.3 (2010), p. 31202. DOI: 10.1117/1.3455989.

[84] Tao Yang et al. "Clustering method for counting passengers getting in a bus with single camera". In: *Optical Engineering* 49.3 (2010), p. 37203. DOI: 10.1117/1.3374439.

[85] Yuedi Yang et al. "Dynamic Origin-Destination Matrix Estimation Based on Urban Rail Transit AFC Data: Deep Optimization Framework with Forward Passing and Backpropagation Techniques". In: *Journal of Advanced Transportation* 2020 (2020). Ed. by Jose E Naranjo, p. 8846715. ISSN: 0197-6729. DOI: 10.1155/2020/8846715.

[86] Ehsan Yavari et al. "Doppler radar sensor for occupancy monitoring". In: *2013 IEEE Topical Conference on Power Amplifiers for Wireless and Radio Applications*. 2013, pp. 145–147. DOI: 10.1109/PAWR.2013.6490217.

[87] Tianwei Yin et al. "Transferable Supervised Learning Model for Public Transport Service Load Estimation". In: (2022). DOI: 10.21203/rs.3.rs-2232957/v1.

[88] Shuo Zhang et al. "Tiny YOLO Optimization Oriented Bus Passenger Object Detection". In: *Chinese Journal of Electronics* 29.1 (2020), pp. 132–138. DOI: 10.1049/cje.2019.11.002.

[89] Jinhua Zhao, Adam Rahbee, and Nigel Wilson. "Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems". In: *Comp.-Aided Civil and Infrastruct. Engineering* 22 (July 2007), pp. 376–387. DOI: 10.1111/j.1467-8667.2007.00494.x.

[90] Juanjuan Zhao et al. "Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems". In: *IEEE Transactions on Intelligent Transportation Systems* 18 (Apr. 2016). DOI: 10.1109/TITS.2016.2587864.

[91] Fang Zhu et al. “Research on Counting Method of Bus Passenger Flow Based on Kinematics of Human Body”. In: *2008 International Conference on Computer Science and Software Engineering*. Vol. 1. 2008, pp. 201–204. DOI: 10.1109/CSSE.2008.611.