UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Análisis del servicio de transporte urbano mediante predicción de la demanda

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Ibáñez Peña, Alejandro

Tutor/a: Julian Inglada, Vicente Javier

Cotutor/a: Jordán Prunera, Jaume Magí

CURSO ACADÉMICO: 2022/2023

Resum

La gestió eficient dels sistemes de transport públic és crucial per a oferir un servei d'alta qualitat als passatgers. Un aspecte important en aquest sentit és la predicció dels nivells de demanda de les diferents línies de transport a fi de tractar d'optimitzar els recursos sense afectar els temps d'espera dels passatgers. D'aquesta manera, en aquest treball es pretén abordar l'anàlisi i disseny dels models adequats per a poder proposar un enfocament nou per a predir els nivells d'ocupació de línies de transport urbà basat en dades històriques, amb la fi última de servir com a entrada per a un simulador de flotes basat en agents. Per a això es planteja l'ús de models basats en tècniques clàssiques i d'aprenentatge automàtic per a predir, amb la major precisió possible, els nivells de demanda i ocupació, així com avaluar el rendiment de l'enfocament proposat utilitzant un conjunt de dades reals d'un sistema de transport urbà.

Paraules clau: Predicció de la demanda, ciutats intel·ligents, flotes urbanes, simulació basada en agents

Resumen

La gestión eficiente de los sistemas de transporte público es crucial para ofrecer un servicio de alta calidad a los pasajeros. Un aspecto importante en este sentido es la predicción de los niveles de demanda de las diferentes líneas de transporte con el objeto de tratar de optimizar los recursos sin afectar a los tiempos de espera de los pasajeros. De esta forma, en este trabajo se pretende abordar el análisis y diseño de los modelos adecuados para poder proponer un enfoque novedoso para predecir los niveles de ocupación de líneas de transporte urbano basado en datos históricos, con el fin último de servir como entrada para un simulador de flotas basado en agentes. Para ello se plantea el uso de modelos basados en técnicas clásicas y de aprendizaje automático para predecir, con la mayor precisión posible, los niveles de demanda y ocupación, así como evaluar el rendimiento del enfoque propuesto utilizando un conjunto de datos reales de un sistema de transporte urbano.

Palabras clave: Predicción de la demanda, ciudades inteligentes, flotas urbanas, simulación basada en agentes

Abstract

Efficient management of public transport systems is crucial to provide a high quality service to passengers. An important aspect in this sense is the prediction of demand levels of the different transport lines in order to try to optimise resources without affecting passenger waiting times. In this way, the aim of this work is to analyse and design the appropriate models to be able to propose a novel approach to predict the occupancy levels of urban transport lines based on historical data, with the ultimate aim of serving as input for an agent-based fleet simulator. This involves the use of models based on classical and machine learning techniques to predict, as accurately as possible, demand and occupancy levels, as well as to evaluate the performance of the proposed approach using a real dataset of an urban transport system.

Key words: Demand prediction, smart cities, urban fleets, agent-based simulation

Contents

Contents v

List of Figures vii

List of Tables vii

1 Introduction 1

1.1 Motivation . . . 1

1.2 Objectives . . . 2

1.3 Expected impact . . . 2

1.4 Memory structure . . . 3

2 State of art 5

2.1 Transport datasets . . . 5

2.2 Related Works . . . 8

2.3 Simulation methodologies . . . 9

2.4 Proposal . . . 10

3 Problem analysis 13

4 Data preparation and comprehension 15

4.1 Data source . . . 15

4.2 Data analysis . . . 15

4.3 Data processing . . . 18

4.3.1 Onboarding-Offboarding processing . . . 19

4.3.2 Resources pre-processing . . . 22

4.4 Post-processing data analysis . . . 24

5 Knowledge extraction and model evaluation 27

5.1 Onboardings prediction . . . 27

5.1.1 Results . . . 28

5.2 Offboardings prediction . . . 30

5.2.1 Results . . . 30

5.3 Resources prediction . . . 32

5.3.1 First version. Deciding the prediction task for each output . . . 34

5.3.2 Second version. Technique experimentation . . . 35

5.3.3 Third version. Combining techniques . . . 36

5.4 Simulation preparation . . . 37

5.4.1 Simulation example . . . 40

6 Conclusions 45

6.1 Future work . . . 46

6.2 Relation of the work developed with the studies pursued . . . 47

Bibliography 49

Appendix

A Project relationship with the Sustainable Development Goals 51

List of Figures

4.1 Passenger traffic magnitude difference of routes 23 Orange, 9 Plum and 14 Peach . . . 16
4.2 Average total passengers per trip for route 2 Green East per weekday and month . . . 16
4.3 Average onboarding passengers per stop for route 1 Red East . . . 17
4.4 Average onboarding passengers per stop for route 11 Cherry . . . 17
4.5 Bus transfers . . . 18
4.6 Ames routes sections division . . . 19
4.7 Average onboarding passengers per section for route 1 Red East . . . 25
4.8 Average onboarding passengers per section for route 11 Cherry . . . 25
4.9 Average number of buses per section for route 1 Red West . . . 26
4.10 Average number of trips per section for route 1 Red West . . . 26

5.1 Boxplot model comparison for route 5 Yellow. Onboarding . . . 30
5.2 Boxplot model comparison for route 14 Peach. Onboarding . . . 31
5.3 Boxplot model comparison for route 23 Orange. Onboarding . . . 31
5.4 Boxplot model comparison for route 5 Yellow. Offboarding . . . 32
5.5 Boxplot model comparison for route 14 Peach. Offboarding . . . 33
5.6 Boxplot model comparison for route 23 Orange. Offboarding . . . 33
5.7 Input example. . . . 40
5.8 Onboarding predictions formatted . . . 41
5.9 Offboarding predictions formatted . . . 41
5.10 Resources predictions formatted . . . 42
5.11 Bus stations JSON load . . . 42
5.12 Buses JSON load . . . 43
5.13 Passengers JSON load . . . 43
5.14 Simulation initial state . . . 43

List of Tables

2.1 Datasets analyzed. . . . 5

4.1 Time periods division. . . . 20
4.2 Results control models. First iteration. Error expressed as MAE . . . 20
4.3 Results control models. Second iteration. . . . 21
4.4 Weighted window for precipitation. . . . 22
4.5 Weighted window for snow depth. . . . 22
4.6 Results control models. Third iteration. . . . 23

4.7 Variable description. Onboarding-Ogboarding dataset. 23
4.8 Variable description. Resources dataset. 24

5.1 Onboarding metrics performance. Routes: (1E: 1 Red East, 1W: 1 Red West, 2E: 2 Green East, 2W: 2 Green West, 3: 3 Blue, 5: 5 Yellow, 7: 7 Purple, 9: 9 Plum, 11: 11 Cherry, 14: 14 Peach, 23: 23 Orange) 29
5.2 Onboarding models assignation per route 30
5.3 Offboarding metrics performance. Routes: (1E: 1 Red East, 1W: 1 Red West, 2E: 2 Green East, 2W: 2 Green West, 3: 3 Blue, 5: 5 Yellow, 7: 7 Purple, 9: 9 Plum, 11: 11 Cherry, 14: 14 Peach, 23: 23 Orange) 31
5.4 Offboardings models assignation per route 33
5.5 Results resources prediction version 1. 35
5.6 Results resources prediction version 2. 36
5.7 Results resources prediction version 3. 37

A.1 Project relationship with the SDGs 52

CHAPTER 1
Introduction

In this chapter, we will describe the motivation that led us to do this project, the goals we imposed and followed to extract the maximum potential for our work, the impact we want to reach and a brief anticipation of the structure of the memory.

1.1 Motivation

Public transportation, particularly buses, plays a critical role in ensuring smart cities are efficient, sustainable, and livable. With growing urbanization, traffic congestion, and environmental concerns, it has become essential to promote the use of public transportation. Buses offer a cost-effective and eco-friendly alternative to private cars, reducing traffic congestion and greenhouse gas emissions. By investing in public transportation infrastructure, smart cities can also promote economic growth, job creation, and social mobility. Therefore, it is crucial to prioritize public transportation as a means of transportation to improve the quality of life for citizens and create more sustainable cities.

Transportation by bus is an essential need in many countries and cities around the world. According to statistics [1], 12% of the population in the United States depend on public transport to commute daily, which is low compared to other countries like South Korea, where it reaches 40% of the total, and it is expected to reach higher levels, making the efficient management of public transportation systems crucial for providing high-quality service to passengers. One important aspect of this is the prediction of bus occupancy levels, as it can help optimize bus routes, improve service reliability, and reduce passenger wait times. Accurately predicting bus occupancy levels can also enable the deployment of dynamic bus scheduling and real-time passenger information systems.

Despite the importance of bus occupancy prediction, there is a lack of research in this area. Most existing studies focus on either real-time sensor data or historical bus occupancy data, which we want to focus on. Historical data let us make long-term prediction studies in different situations in the process of "whatification" and detect possible problems in the public transportation system, such as bottlenecks, shortages during rush hours, or wastage of resources, among others, which lets prevent those problems by enacting modifications over the route planning or resource assignment.

[1] https://www.statista.com/chart/25129/gcs-how-the-world-commutes/

1.2 Objectives

The aim of this project is to build passenger demand and resource assignment prediction models to serve as the input of simulation tools.

In order to make it an easier task, it is divided into the following goals:

- Analyze the publicly available datasets on public urban transportation and study the feasibility of each one to perform the project to reach a final one.
- Research previous works based on similar ideas regardless of being bus system-oriented or any other public transport and extract key insights that could be used in our study.
- Analyze different simulation techniques and choose the best approach based on our requirements.
- Study the current situation of public transport, its weaknesses and points that could be improved.
- Analyze the chosen dataset to characterize the data and extract ideas that can affect posterior processing and modelling.
- Process the data to get a proper format for modelling.
- Create a model for each of the three predicting tasks: passenger onboardings and offboarding, and resources assignation.
- Creating a pipeline to format the predictions into an interpretable format suitable for the simulation.

1.3 Expected impact

With this project, we want to take a step forward towards a fully sustainable and efficient urban transport system. In response to the growing demand for strategies to address passenger traffic demand, we aim to contribute to that mission by employing a data-based approach. With this, we can gain a comprehensive understanding of passenger traffic patterns, over and underused resources and general travel preferences. These insights and methodology derived from this research can serve as inspiration and guidance for policymakers and decision-makers of transportation institutions, who could use this methodology for their own data-based transportation solutions. Therefore, the appliance of an optimized resource allocation in transport systems can result in enhanced service quality.

Having an optimized transport system is undeniably related to a reduction of environmental impact. By improving resource allocation, our research project can contribute to the reduction of greenhouse gas emissions. The enhancement of service quality leads to higher reliability of public urban transportation methods, which translates into an increased use of the service. This can significantly reduce the use of private vehicles, which results in lower carbon emissions.

Furthermore, transportation institutions can benefit from the economic advantages of employing data-driven strategies. The optimization of bus routes and schedules can minimize inefficiencies, which always translate into unnecessary excesses in fuel consumption, maintenance costs and operational expenses. As resources can be then utilized in the most cost-effective manner, institutions can allocate resources to other critical areas of system development and service enhancement.

1.4 Memory structure

The memory is structured as follows:

- **Chapter 1, Introduction**: In this chapter, we talk about the motivation, objectives, expected impact, and structure of the report itself.
- **Chapter 2, State of art**: Here, we review public data available about urban transportation systems, related work, and simulation techniques. We also give an overall description of the methodology that will be developed in the project.
- **Chapter 3, Problem analysis**: In this chapter, we study the weaknesses of urban transport, as well as possible improvements.
- **Chapter 4, Data preparation and comprehension**: Here, we describe the dataset chosen with an analysis of the main features. The data is then processed and formatted in a proper format for model building.
- **Chapter 5, Knowledge extraction and model evaluation**: In this chapter, we build the models for onboardings, offboardings and resource allocation prediction. Additionally, a pipeline is built to format the predictions to serve as the input of a transportation simulation tool.
- **Chapter 6, Conclusions**: In this last chapter, we detail the objectives achieved. Additionally, we talk about the relationship of this project with the Data Science Degree undertaken. Finally, we discuss how this project could be improved with further research.

CHAPTER 2
State of art

In this chapter, we will revise some of the most important public transport datasets, analyzing the advantages and disadvantages, and the general feasibility of doing a project based on each one. Following is a look at the most influential works on passenger flow prediction in different modes of transport, synthesizing the crucial ideas and how each one inspired us to use some of these methodologies in our project. Then we will talk about the different transport simulation methodologies with a brief description of each one. Finally, we are going to introduce the solution proposed in this project.

2.1 Transport datasets

Table 2.1: Datasets analyzed.

Name	Description	Source	Time extent	Features
Syd	Train, bus, ferry and light rail services of Sydney, the Blue Mountains, Central Coast, Hunter, Illawarra and Southern Highlands. The granularity is at a stop level. Occupancy_status could be the target variable. Categorical with three states whether the occupancy is over the seating capacity, less than half of the seating capacity or other.	Transport for NSW [10]	08/08/2016 - 14/08/2016 21/11/2016 - 27/11/2016 26/12/2016 - 01/01/2017	Date, Route, Direction, Hour, Trip_id, Bus_id, Stop, Suburb, Latitude, Longitude, Capacity, Seating_capacity, Occupancy_status

Continued on next page

Table 2.1: Datasets analyzed. (Continued)

Tor	Bus services of Canada. Not a dataset per se but an in-live bus occupancy application that could be scrapped. Occupancy_status could be the target variable. Categorical with three states depending on the number of passengers: 0-14, 15-40, >40.	Toronto Transit Commission [6]	Live	Date, Route, Hour, Stop, Latitude, Longitude, Occupancy_status
Tra	Any location whose local agency is a client of TransLoc. Not a dataset per se, but an API to create the dataset. The API has severe request limitations.	TransLoc [23]	Live or historical	Date, Route, Hour, Stop, Latitude, Longitude, Capacity, Occupancy
TFL	Underground service of London. Not a dataset per se, but an API to create the dataset. The occupancy could be inferred using the delay.	Transport for London [9]	Live	Date, Route, Hour, Stop, Delay, Latitude, Longitude
Sao	Bus service of São Paulo. Two datasets. Historical at a daily aggregation level. Real-time at stop level. The variable 'Total_occupancy' can be used to estimate the occupation at each stop based on the delay.	SP Trans of São Paulo [3]	Historical: Jan 2013 - Apr 2020 Real-time: 28/04/2020 - 06/05/2020	Date, Route, Hour, Stop, Delay, Latitude, Longitude, Occupancy, Total_occupancy

Continued on next page

Table 2.1: Datasets analyzed. (Continued)

Joh	Bus service in Johannesburg. The granularity is at a stop level. Three routes and a bus for each.	Johannesburg public transportation [18]	25/02/2019 - 01/03/2019	Date, Route, Hour, Stop, Latitude, Longitude, Onboardings, Offboardings, Speed
Ger	Bus service in a medium-sized German city (not specified). The granularity is at a stop level. Four routes and a bus for each.	[16]	01/01/2022	Date, Route, Hour, Stop, Occupancy
Ame	Bus service of Ames, Iowa. The granularity is at a stop level.	Ames Transit Agency and CyRide [25]	Oct 2021 - June 2022	Date, Route, Hour, Trip_id, Bus_id, Stop, Capacity, Onboardings, Offboardings, Delay
Seo	Bus service in Seoul. The granularity is at a stop level. All the text is in Korean.	Seoul Bus [1]	May 2022	Date, Route, Hour, Stop, Onboardings, Offboardings

Table 2.1 shows the datasets reviewed for the analysis, with information about the time extent, variables included, and a brief description of the most notable aspects of each one. We will now discuss the criteria used to discard the datasets until reaching a final one, which will be used for the project.

Time extent Since we want to build models that capture passenger flow dynamics across different seasons, we need the data to cover a wide time expanse. Therefore we will not consider datasets with less than three months of data. With this, the datasets 'Syd', the real-time one from 'Sao', 'Joh', 'Ger' and 'Seo' are discarded. It is worth noting that, for the moment, the datasets that require the creation of the dataset itself, the ones tagged as live, are kept, since we could spend a minimum of three months collecting the data.

Demand prediction feasibility To predict the demand at a stop level, we need an accurate value or range of values. Therefore, any dataset that implies inferring those values is not trustworthy at all, so we prefer to discard them. This is the case with the datasets 'TFL' and the historical one from 'Sao'. Further on this point, we will also discard 'Tor', since we can not get the onboardings and offboardings at each stop, but a change in the occupancy status; as this is not remarkably precise, we prefer not to take it into account.

With this, we only have two datasets left, 'Tra' and 'Ame'. Both datasets have similar variables, have an extensive time span and are suitable for model prediction. Nevertheless, if we choose 'Tra', we will need to extract the data via the API, having to deal with the request limitations, which will take an undetermined amount of time to have a proper dataset. Therefore, as 'Ame' is already a legitimate dataset with the requirements we imposed, we will go further on working with its data.

2.2 Related Works

Bus passenger flow prediction has been a widely discussed topic in recent years. We can observe two categorizations of the approaches, depending on the time scope, real-time or atemporal prediction, and models themselves, traditional or deep learning methods, where the combination of real-time and deep learning is the clear favourite.

In the line of this last group, we can find [2], which redefines the concept of a bus network as a graph, where the nodes represent the bus stops, while the edges, the segments of the routes, which are directed. As stops are organized in different routes, it is necessary to make a distinction between routes and about the order of the stops in the route; in this case, an adjacency matrix is built, where any pair of nodes where the first one is the immediate predecessor of the second one is assigned a 1, and 0 otherwise. Furthermore, as stops have an influence on other stops, the closer, the stronger; this information is represented in a proximity matrix, in which any pair of nodes of the same route is assigned a 1 if the second stop is reachable from the first one under a predefined t time, and 0 otherwise. The model then consists of a graph-based convolution operator layer, which uses the adjacency and proximity matrices to dictate the relationship of the stops between them, combined with a Long Short-Term Memory (LSTM) model for short-term passenger prediction. This novel approach shows accurate results for short-term prediction tasks. Nevertheless, the study lacks external data that can affect passenger flow, such as festivities or environmental conditions, among others.

In [19], a combination of three deep learning techniques is proposed. Firstly, a greedy-layer algorithm is used to prepare the data for prediction by fetching the numerical columns from the original dataset and relating it to each region; then, an LSTM layer filters the data batches by removing redundant data, which are the input of a Recurrent Neural Network (RNN), that produces a final prediction. This approach achieves highly positive results.

For the case of a real-time forecasting system, in these two papers, we could observe that the LSTM technique is widely used, so treating our data as a series may be a suitable solution for our purpose.

Nevertheless, we are going to have a view into atemporal prediction approaches. In this area, we found [15], which uses two different networks based on an LSTM layer, in one case uni-directional and bi-directional in the other, for passenger prediction within a time range. This inspired us to use an LSTM-based network, with the difference of just employing uni-directional LSTM layers; since the bi-directional layers are not suitable for a real-world environment; as for a prediction in t+1, the time-segments t+2 and following are used as inputs.

[17] proposes a multistep-ahead prediction hybrid system using an ARIMA-SVM model. The model consists of two stages, where an Autoregressive Integrated Moving Average (ARIMA) model is used to analyze the linear part of the problem, and a Support Vector Machine (SVM) model is developed to model the residuals from the ARIMA model, which contains information about the non-linearity. Two multistep-ahead forecasting strategies are employed, iterated, which predicts one sample at a time, and direct, which takes the dataset in groups of n sequential samples to be predicted in one go. We

found it interesting to use an SVM technique applied for regression to fit non-linear data, as after the aggregation, the data gains non-linearity.

[26] uses a Grey Prediction Model (GM), specifically the first order Grey Model GM(1,1), as, due to environment complexity, it is commonly used a single factor, which is the series itself. The main benefit of this model is that it distinguishes patterns within the days of the week and from week to week. That is, a Monday in the week t+2 will be mainly influenced by the Mondays in weeks t+1, t, and so, but also by the trend observed in the whole week t+1. As the data itself can not have enough autocorrelation, plus it can be fed with other variables, a GM(1,1) model may not be the most suitable approach; nevertheless, this inspired us to use a variant of it, the GM(1,n). This model makes use of the serial variables to make the prediction to supply the lack of self-explicability.

2.3 Simulation methodologies

A bus system is a sophisticated and dynamic transportation system that must deal with a variety of unpredictable factors, including weather conditions, traffic jams, accidents, and passenger demand. It could be required to rearrange the bus system in order to maintain service levels, limit interruptions, and maximise resource utilisation when unplanned occurrences, such as road closures or delays, occur. Bus system rescheduling is a challenging issue that needs thorough design, research, and implementation.

Historically, manual techniques like spreadsheet-based scheduling or manual dispatching have been used for bus system rescheduling. These procedures take a significant amount of time, are labour-intensive, and are error-prone. Additionally, they might not be able to react to rapid changes in demand or unforeseen occurrences like road accidents or bad weather.

Simulation-based approaches have emerged as a promising alternative to manual methods, allowing researchers to model the bus system's behaviour, evaluate different rescheduling strategies, and assess their performance under various scenarios. Simulation models can depict the intricate interactions between the many elements of the bus system, such as the buses, passengers, and infrastructure, more precisely and realistically.

Simulation-based techniques for bus system rescheduling can be broadly categorized into three categories: microscopic, mesoscopic, and macroscopic.

Microspcopic models Microscopic models simulate the behaviour of individual buses, passengers, and infrastructure components in real time. These models are quite accurate and may depict the intricate relationships between various bus system components. However, they may not be appropriate for a larger-scale analysis of the behaviour of the bus system since they demand a significant amount of data and computation power.

Mesoscopic models Mesoscopic models simulate the behaviour of buses and passengers at an aggregate level, focusing on the overall flow of passengers and buses through the system. These models are appropriate for analysing the behaviour of the bus system on a broader scale thanks to their ability to find a compromise between accuracy and computing complexity. However, they might not be able to record the intricate interactions between specific bus system parts.

Macroscopic models Macroscopic models simulate the behaviour of the bus system at a high level, focusing on the overall performance metrics, such as travel time or passenger satisfaction. These models are simple and easy to use, but they might not be able to reflect the intricate relationships between the many parts of the bus system.

In recent years, several research studies have proposed novel approaches for bus system rescheduling based on simulations. Here are some recent advances in this field:

1. Multi-objective optimization: Some studies have proposed multi-objective optimization approaches that can simultaneously optimize several performance metrics, such as service level and operational costs. These approaches can identify trade-offs between different objectives and provide a more comprehensive evaluation of rescheduling strategies.

2. Real-time rescheduling: Real-time rescheduling approaches aim to adjust the bus schedules on the fly in response to real-time events, such as traffic congestion or passenger demand. These approaches can improve the responsiveness of the bus system and reduce delays and wait times for passengers.

3. Machine learning-based approaches: Some studies have explored the use of machine learning algorithms to predict bus demand and optimize bus schedules. These approaches can improve the accuracy of demand forecasting and lead to more effective rescheduling strategies.

4. Agent-based models [12]: Each bus or passenger is represented as an agent interacting with the other agents in the system. This method is capable of capturing the bus system's dynamic behaviour as well as the interactions among the vehicles, people, and surroundings. Agent-based models, however, may be computationally demanding and need substantial data and parameter adjustment.

5. Discrete event simulation [11]: The bus system's activity is modelled as a series of events that take place across time. Each event represents a change in the bus system's status, such as a bus pulling up to a stop, a passenger getting on or off, or a traffic delay. Although discrete event simulation can handle large-scale systems more effectively than agent-based models, it could not adequately depict the dynamic behaviour of the bus system.

6. Hybrid simulation models: Hybrid simulation models combine different simulation techniques, such as microscopic and macroscopic models, to capture the behaviour of the bus system at different scales. These models can provide a more comprehensive and accurate representation of the bus system's behaviour and help identify the most effective rescheduling strategies.

2.4 Proposal

Everything said now, we are going to define in general terms the solution that will be developed in the following chapters. We will use the dataset 'Ame' (see Table 2.1) since it covers a large time span, has a stop level granularity, and, in general, the format is suitable for building proper data models. About the approach, we will focus on atemporal prediction, since the objective of this project is not to be deployed into a real-world system, which means real-time predictions, but for simulations, that are completely atemporal. As in this approach we can use both traditional and deep learning methods, we will create models based on both methodologies. Finally, we will target our models to produce outcomes suitable for microscopic, agent-based simulations, since that is the format required by the simulation tool we will be using, called SimFleet [20]. Further on this tool will be explained in Chapter 5.4.

In the next chapter, we will discuss the problems of resource allocation in transport and the importance of good demand forecasting.

CHAPTER 3
Problem analysis

The urban transport system plays an undeniably vital role in the functioning and development of cities, affecting the quality of life for residents, the efficiency of businesses, and the overall sustainability of urban areas. As cities continue to grow and travel patterns evolve, providing efficient and reliable urban transport services becomes then an increasingly challenging task. Accurate demand forecasting is a critical component of effective urban transport planning, as it enables policymakers, transport authorities, and service providers to anticipate future travel patterns, allocate resources efficiently, and design transportation systems that meet the needs of a rapidly changing urban landscape.

One significant challenge in addressing the problems associated with the resource management of bus networks is the limited availability of public transportation datasets. Although there are some publicly accessible datasets related to transportation, they often lack the proper granularity and specificity required for accurate demand forecasting and resource optimization. The scarcity of comprehensive and up-to-date datasets hinders the development and evaluation of effective models and solutions.

Further on this issue, publicly available datasets normally lack a standard and consistent format. Datasets may vary in terms of the variables recorded, the time intervals at which data is collected, and the geographical coverage. This lack of uniformity makes it challenging to compare and integrate datasets from different sources, limiting the ability to derive meaningful insights or build robust models. Additionally, incomplete or inconsistent data can lead to biased or inaccurate results when developing models for resource management.

The absence of comprehensive and standardized datasets poses significant implications for developing models that can effectively address the challenges associated with resource mismanagement in bus networks. Without access to reliable and detailed data, it becomes difficult to capture the complexity and dynamics of urban transportation systems accurately. Consequently, the performance of models and solutions may be suboptimal, limiting their potential to mitigate issues such as congestion, inefficient routing, and inadequate service levels.

Collaboration and data sharing among various partners become crucial to overcome the limitations imposed by the lack of public transportation datasets. Transportation authorities, service providers, and researchers should have a key role in collecting, consolidating, and sharing relevant data in a standardized format. These open data initiatives can play a vital role in enabling the development of more accurate and comprehensive models for resource management in bus networks.

Data collaboration is undeniably essential; nevertheless, it is a must to address privacy and ethical concerns associated with transportation datasets. To safeguard personal information and ensure compliance with data protection regulations, it is required the implementation of anonymization and appropriate data governance practices. A balance of data openness and privacy is crucial to encourage collaboration while maintaining public trust and protecting individual privacy rights.

To sum up, we want this project to motivate public institutions to move towards more public and collaborative transportation data sharing to achieve creating a fully reliable and efficient bus network system.

In the next chapter, we will analyze the dataset we selected and process it to achieve the desired format to proceed with the model building.

CHAPTER 4

Data preparation and comprehension

In this chapter, we are going to describe the dataset and deeply analyze it, alongside the processing procured for the accomplishment of our objectives.

4.1 Data source

The bus occupancy data were obtained from multiple routes operated by Ames Transit Agency in Ames, Iowa, metropolitan area from October 2021 to June 2022 [25]. The data is divided into 9 CSV files, one per month, which consist of a total of 4,577,930 records. The data was collected through a system of automatic passenger counters (APCs) so that each database entry corresponds to the passenger boarding and alighting of each bus at each stop. In addition to this information, each individual details the route, the scheduled and actual arrival and departure times, and the capacity of the vehicle. There are a total of 9 circular routes, 3 linear routes; where the outward and inward journeys are marked as two different routes and 1 hybrid route; which during the school year works as a linear route and as a circular route during summer. Of these routes, 4 show operational inconsistencies: 1 changes part of the journey on weekdays and weekends, 1 has limited service, causing periods of lack of data, and 2 do not operate or change the part of the journey during summer. There are a total of 5 bus types, depending on the capacity of passengers, which from an order of usage are 60 (70.02%), 65 (23.16%), 90 (4.84%), 40 (1.63%), and 20 (0.35%).

4.2 Data analysis

Before getting hands-on preprocessing the data, it is important to analyze the data itself. This is a crucial step in understanding the nature and characteristics of the data structure of each route in the Ames bus system. Furthermore, this process can help ensure that the preprocessing steps are appropriate and effective by being guided on the basis of patterns and relationships found in the data.

For this process, only the routes without inconsistencies will be taken into account. These routes are 1 Red East and 1 Red West (linear routes), 2 Green East and 2 Green West (linear routes), 3 Blue, 5 Yellow, 7 Purple, 9 Plum, 11 Cherry, 14 Peach, and 23 Orange.

Our first consideration was whether the passenger load was evenly distributed throughout the bus system, as well as throughout the year, in other words, if we would find similar passenger traffic values regardless of the route and month. A sample of this study can be seen in Figure 4.1, where we compare the average total onboardings during a single trip for the routes 23 Orange, 9 Plum and 14 Peach along all the months we have data of. As we can see, an average trip on the route 23 Orange carries, in most cases, almost twice as people as in 9 Plum and more than 4 times as in 14 Peach; which shows a clear disbalance in the demand of the routes of Ames bus system. Furthermore, we can see a clear pattern of traffic decrease during festivities or no class period, as during the months of December, March, May and June; which to Christmas, Spring Break and the end of the exam period for the last two months in our data. This gives us a key clue about the main type of users of the transport system, which are the students at Iowa State University.

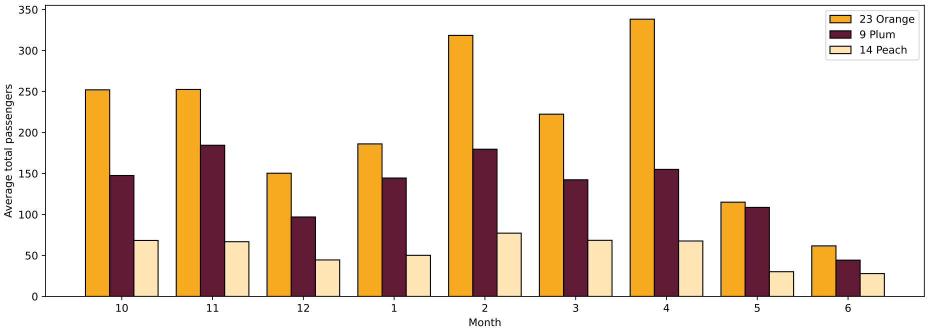


Figure 4.1: Passenger traffic magnitude difference of routes 23 Orange, 9 Plum and 14 Peach

As public transport traffic has shown to follow periodic weekly trends [14], we wanted to observe the presence of that effect in our data and if it is strong enough to be significant for the upcoming modelling. Figure 4.2, shows this effect for route 2 Green East, where there seems to be a trend of traffic decrease over the week until the weekend, where this decline is abruptly accentuated. This effect is similar for the rest of the routes, except for 5 Yellow, which does not have service the Sundays, and routes 7 Purple, 9 Plum, 11 Cherry, 14 Peach, and 23 Orange, which neither have both for Saturdays and Sundays.

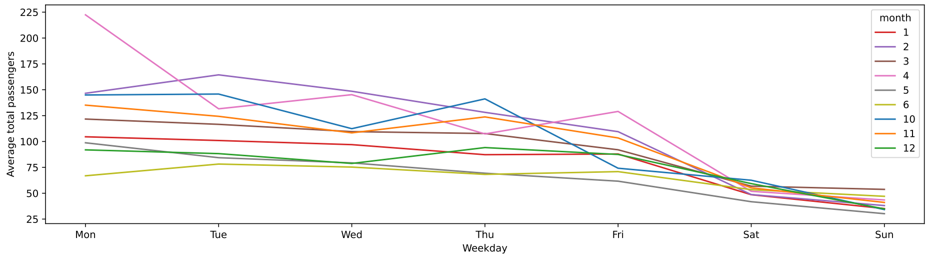


Figure 4.2: Average total passengers per trip for route 2 Green East per weekday and month

We have already analyzed how the traffic is distributed along the routes and both along the week and year, so now we want to focus on a lower level of aggregation, as is the distribution over all the stops of a route during the day. Figure 4.3 represents this distribution for route 1 Red East, where we can clearly see two popular onboarding zones, as are the stops corresponding to the first group of stops at Lincoln Street and the ones nearby, and the second group of stops at Lincoln Street and the ones immediately prior to them. Furthermore, we can identify each one with a moment of the day, being the first case from the morning until somewhere around midday; while the second one is clearly more popular during the afternoon and evening. This morning-afternoon difference is more notable in Figure 4.4, which represents the same information as the previous one, but for route 11 Cherry. During the morning, the most popular stops to get on the morning are located left side of the image, whereas the inverse occurs on the other side, where the popularity rises as we approach late hours. Similar behaviours are present in the rest of the routes. This shows us two key things, that stops have similar behaviour to their neighbours, and so happens with the hours, there is no abrupt change in the traffic from hour to hour.

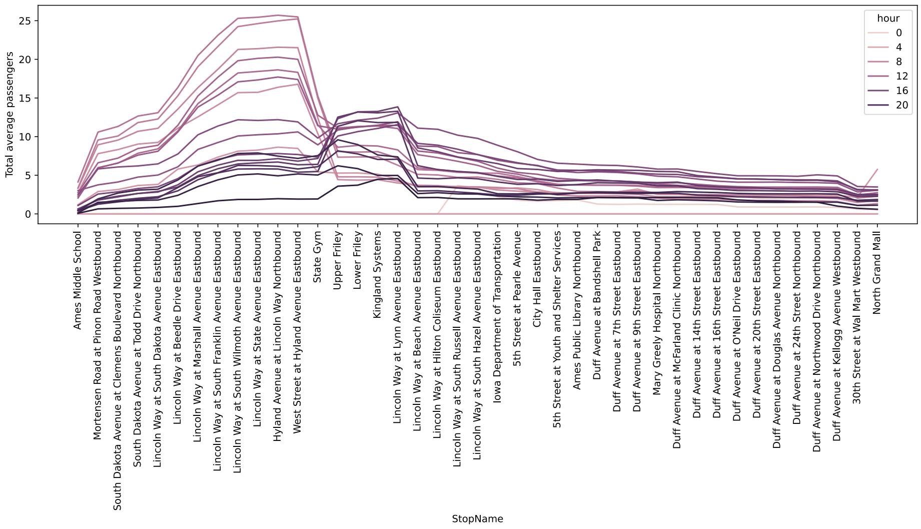


Figure 4.3: Average onboarding passengers per stop for route 1 Red East

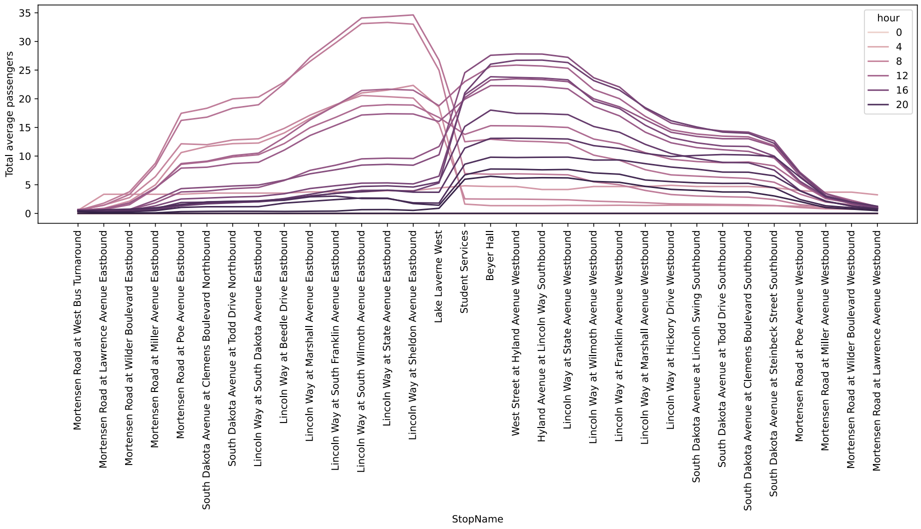


Figure 4.4: Average onboarding passengers per stop for route 11 Cherry

As there are two couples of linear routes, it is interesting to study if it is possible to circularize them. Figure 4.5 shows the number of buses that transfer from an initially assigned route to another one. As expected, the pairs of routes that have the most transfers between them are 1 Red and 2 Green couples of linear routes; nevertheless, the number of transfers does not fit the total quantity of buses assigned, which means that, when a bus finishes the trip in one direction, not always returns to make the complementary one. From West to East, the transfers are 1,530 for 1 Red and 990 for 2 Green, whereas the number of buses assigned is 1,816 and 1,036, respectively. With this, it is clear that circularizing these routes can not be an option, so they will remain as independent linear routes.

Furthermore, there are also transfers between pairs of routes outside the before-mentioned relationship, which means that the bus system responds to demand requirements by modifying the way its resources are allocated, which has to be taken into account in the simulation setup.

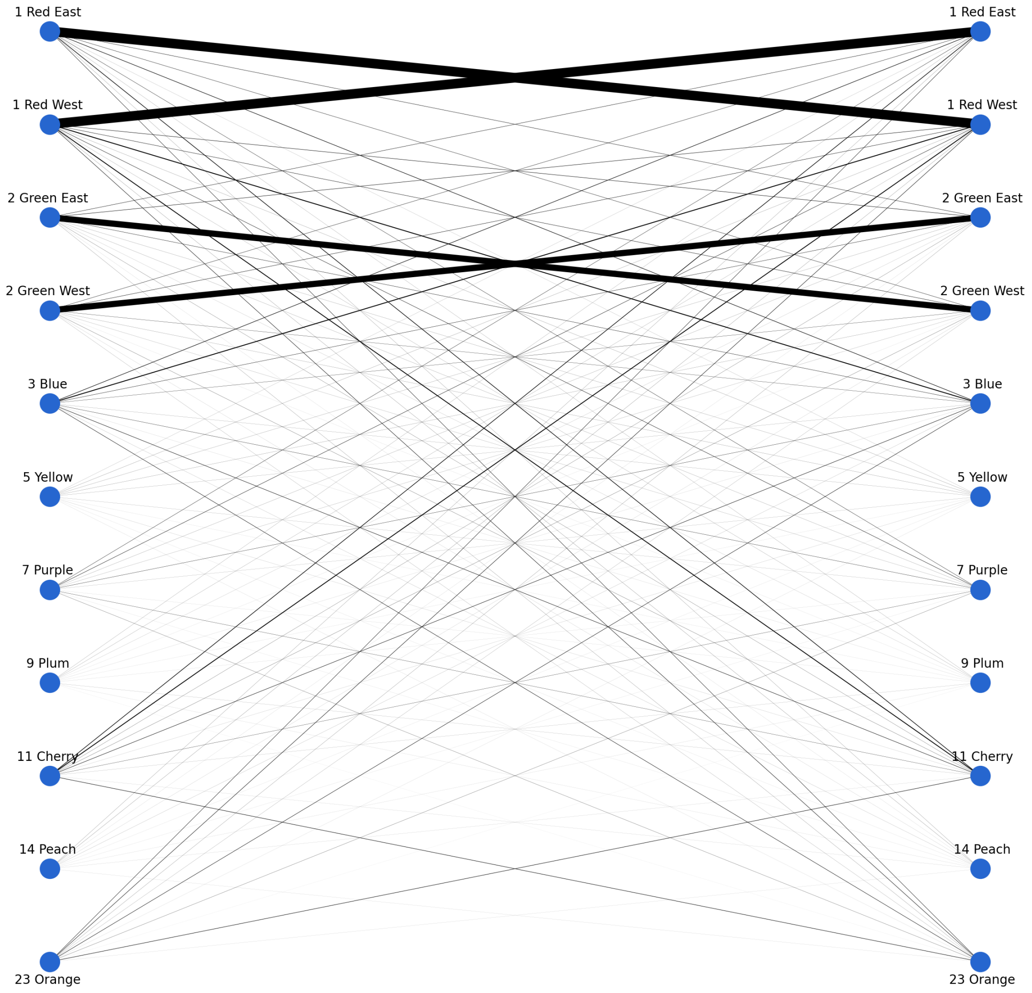


Figure 4.5: Bus transfers

4.3 Data processing

The dataset was first reduced by keeping the routes that did not show inconsistencies, having then 2 linear and 7 circular routes; linear routes have not been transformed into circular ones since the main requirement that having an outward journey means having a return one was rarely met, as seen in Chapter 4.2. Furthermore, during the period of data collection, the routes have undergone variations due to the elimination or addition of stops, for which the dataset has been adapted by eliminating the stops that no longer exist and ignoring the new ones. Further research showed that some routes had more than one pattern, as due to certain events, the route is temporarily altered; nevertheless, due to the timeliness of these cases, it was decided to omit them from the dataset.

Since the project aims to estimate the number of passengers that need the bus service around an area within a certain time range and the data is presented as records, they have to be aggregated to create flows. For this process, the stops will be grouped into sections, and the time ranges must be defined. The criteria followed to determine the sections within a route was, in the first instance, dividing each route taking the neighbourhoods of Ames as a basis and then subdividing those groups considering changes in the surroundings of a bus stop, such as transitioning from a residential to a commercial zone; always taking a minimum of three stops per section. The exact section division can be seen in Figure 4.6. Regarding the time zones, the schedule of Iowa State University[1] was followed as a basis, creating 7 time periods to address the fluctuation of passenger occupancy. The time division can be seen in Table 4.1, where the time ranges night periods of the previous and current day have been aggregated into the '1_night' period since they represent, in fact, the same period. The aggregation is then done having the route as the first level, then the section, followed by the date and finally, the time range.

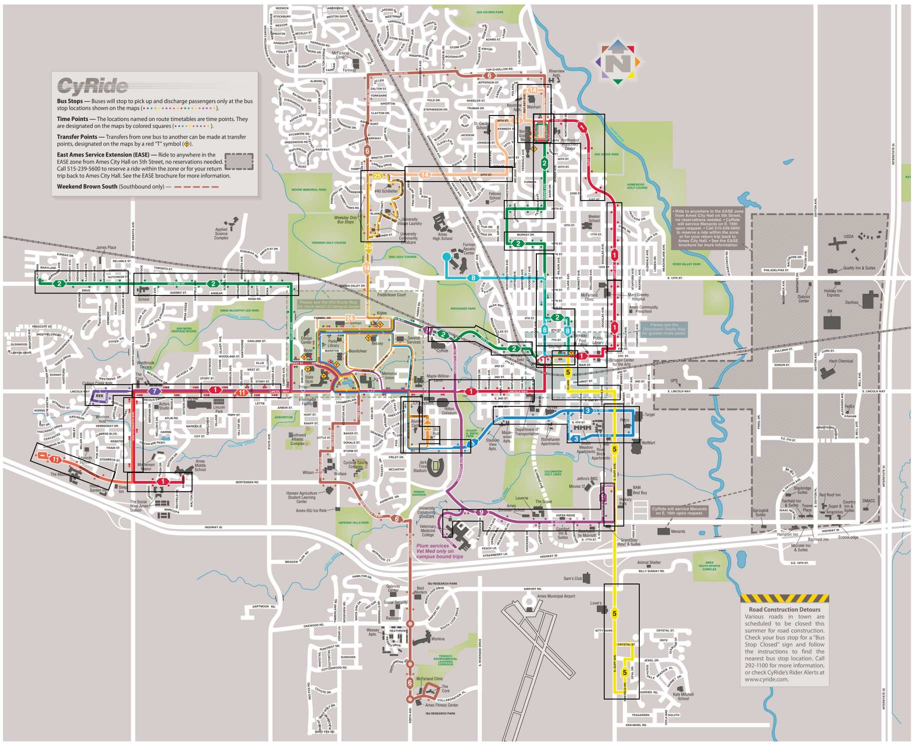

CyRide
Bus Stops — Buses will stop to pick up and discharge passengers only at the bus stop locations shown on the maps (• • • • • • • • • • • • • • • • • •).
Time Points — The locations named on route timetables are time points. They are designated on the maps by colored squares (▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪).
Transfer Points — Transfers from one bus to another can be made at transfer points, designated on the maps by a red "T" symbol (◆).
East Ames Service Extension (EASE) — Ride to anywhere in the EASE zone from Ames City Hall on 5th Street, no reservations needed. Call 515-239-5600 to reserve a ride within the zone or for your return trip back to Ames City Hall. See the EASE brochure for more information.
Weekend Brown South (Southbound only) — — — — — —
Road Construction Detours
Various roads in town are scheduled to be closed this summer for road construction. Check your bus stop for a "Bus Stop Closed" sign and follow the instructions to find the nearest bus stop location. Call 292-1100 for more information, or check CyRide's Rider Alerts at www.cyride.com.
N

Figure 4.6: Ames routes sections division

4.3.1. Onboarding-Offboarding processing

The aim of this part of the project is to give correct predictions of the passenger demand for the current bus system, this will help us create realistic scenarios of passenger traffic during the simulations.

A series of iterations were performed to analyze if the data quality was good enough to develop more complex models, which both will give better results and consume more

[1] https://www.event.iastate.edu/?sy=2023&sm=01&sd=26&featured=1&s=d

Time period	Starts	Ends
1_night	22:00 (previous day)	4:59
2_early_morning	5:00	8:59
3_morning	9:00	11:59
4_midday	12:00	14:59
5_afternoon	15:00	18:59
6_evening	19:00	21:59

Table 4.1: Time periods division.

Route	RF	SVR
1 Red East	6.252	6.443
1 Red West	5.931	6.861
2 Green East	3.080	3.292
2 Green West	2.674	3.146
3 Blue	22.908	24.245
5 Yellow	2.059	1.954
7 Purple	4.934	5.482
9 Plum	20.399	18.291
11 Cherry	8.370	10.653
14 Peach	1.298	1.216
23 Orange	162.555	145.245
General	**9.958**	**10.053**

Table 4.2: Results control models. First iteration. Error expressed as MAE

time to process. To control the quality of these approaches, we will use two simple models as are the RF and SVR models, over the prediction of the total onboardings of each time and zone segment. In order to compare the quality improvement of each iteration, we will use the mean absolute error metric (MAE). It is worth mentioning that each route has its own independent model trained and tested exclusively with its data.

First version: Bus data

The first iteration was a simple contact point, where we wanted to observe the predictability of the data after the pre-processing developed in the previous section 4.3.

As seen in Table 4.2, the data seems correct to be used in prediction, as the errors in most of the routes are within an acceptable range according to the capacity of the most frequent buses (mentioned in section 4.1), with the exception of routes 3 Blue, 9 Plum, 11 Cherry and 23 Orange.

Second version: Enrichment with meteorological data

In this second iteration, we wanted to give some context about the environmental conditions of each time record, as it has some significant effect on the behaviour of passenger traffic in public transport [24].

Route	RF		SVR	
	Model MAE	Improvement	Model MAE	Improvement
1 Red East	5.678	0.574 (9%)	6.345	0.098 (2%)
1 Red West	5.378	0.553 (9%)	7.066	-0.205 (-3%)
2 Green East	2.842	0.238 (8%)	3.282	0.010 (0%)
2 Green West	2.353	0.321 (12%)	3.173	-1.127 (-1%)
3 Blue	19.109	3.799 (17%)	22.104	2.141 (9%)
5 Yellow	2.054	0.005 (0%)	1.958	-0.004 (0%)
7 Purple	4.407	0.527 (11%)	5.360	0.122 (2%)
9 Plum	15.840	4.559 (22%)	15.927	2.364 (13%)
11 Cherry	8.253	0.117 (1%)	11.873	-1.220 (-11%)
14 Peach	1.240	0.058 (4%)	1.206	0.010 (1%)
23 Orange	117.852	44.703 (28%)	128.256	16.989 (12%)
General	8.196	1.762 (18%)	9.529	0.524 (5%)

Table 4.3: Results control models. Second iteration.

To enrich the quality of the data by studying the effect of different meteorological conditions, a weather dataset obtained from the Iowa Environmental Mesonet[2] was appended. The selected station was the AMES 5 SE, with code IA0203. As the data did not follow the same granularity as the already processed data, instead it had daily information. At the first moment, it was decided that, regardless of the hour range, the daily value would be kept for all the individuals within that day. From this new dataset, we only considered significant to have an effect on passenger traffic data about precipitation, which includes a variable for rain + melted snow and another for snow; and its remains, which is the value of inches of accumulated snow.

Table 4.3 shows that there is an improvement in the results of all the routes for the RF model when compared to the first iteration, where the error decrease is around an 18% overall, mainly fed by the routes that showed the highest errors that here are the most benefited ones. Although SVR also shows some improvement, around a 5% overall, not all routes are benefited. Nevertheless, the routes that have a higher error in this iteration than in the previous one have an increment of around or less than 0.2, with the exception of route 11 Cherry, which is over 1.2. In the remaining routes, the ones which have benefited, the improvement has not been as notable as in the RF.

Third version: Improving the data quality

In this last iteration, we wanted to refine the quality of the current data by improving the format of the dataset.

As the occupancy rate in public transportation follows weekly trends [14], as seen in Figure 4.2, two additional columns were created to keep that context: "ons_yesterday" and "ons_last_week". "ons_yesterday" stores the total onboardings during the previous day at the same route, section and hour range, whereas "ons_last_week" does the same but for the previous week at the same weekday. With this, a general trend is procured, and a particular trend for each weekday. It is worth noting that, these columns are only

[2] https://mesonet.agron.iastate.edu/request/coop/fe.phtml

Time period	Previous day	Current day
1_night	5/6	1/6
2_early_morning	4/6	2/6
3_morning	3/6	3/6
4_midday	2/6	4/6
5_afternoon	1/6	5/6
6_evening	0/6	6/6

Table 4.4: Weighted window for precipitation.

Time period	Current day	Following day
1_night	6/6	0/6
2_early_morning	5/6	1/6
3_morning	4/6	2/6
4_midday	3/6	3/6
5_afternoon	2/6	4/6
6_evening	1/6	5/6

Table 4.5: Weighted window for snow depth.

used when predicting the onboarding, the same pair but for offboardings were also created although, of course, are not used for comparing the improvement of this iteration.

As the weather data is not uniform during the day, we wanted to give a logical value to each individual. Rather than keeping the daily value regardless of the hour range or a sixth of it, it would be better to infer the value in each hour range by considering the next and previous days as a moving window. For the precipitation variables (rain + melted snow, and snow), as the daily value represents the cumulative precipitation for the whole day, the inferred hour range value would be a sixth of the weighted total from the previous and current day; the weight proportions can be seen in Table 4.4, as it is considered that this method would consider the precipitation trends and attenuate peak downpours. Nevertheless, as the snow depth value is taken first thing in the morning, it would be considered the melting of the snow by having a weighted average of the current and following day; following the proportions shown in Table 4.5.

As seen in Table 4.6, the refining performed in this last iteration has resulted in an enormous improvement in every route, which results in an overall improvement of 24% for RF, while for SVR, it reaches a 29%. Route 23 Orange shows the highest error decrease, with a difference of over 45 passengers in the MAE for the SVR model, and over 41 for RF, a 36% decrease in both cases, closely followed by route 3 Blue, with improvements of almost 8 passengers for SVR and almost 6 for RF, which represents a 36% and 30% improvement, respectively. The only route that shows an error increase is 5 Yellow in the SVR, although it is not significant, as the increase is only of 0.02 passengers in the MAE metric, or a -1% worsening.

The resultant dataset that will be used for prediction has a final size of 79,741 records. A brief description of each field can be seen in Table 4.7.

4.3.2. Resources pre-processing

The aim of this section is to accurately predict the resources that will be employed in different scenarios created during the simulations. We have two different types of re-

Route	RF		SVR	
	Model MAE	Improvement	Model MAE	Improvement
1 Red East	5.030	0.648 (11%)	5.034	1.311 (21%)
1 Red West	4.793	0.585 (11%)	5.418	1.648 (23%)
2 Green East	2.326	0.516 (18%)	2.425	0.857 (26%)
2 Green West	2.124	0.229 (10%)	2.311	0.862 (27%)
3 Blue	13.322	5.787 (30%)	14.185	7.919 (36%)
5 Yellow	1.980	0.074 (4%)	1.978	-0.020 (-1%)
7 Purple	3.859	0.548 (12%)	4.020	1.340 (25%)
9 Plum	12.064	3.776 (24%)	13.481	2.446 (15%)
11 Cherry	7.191	1.062 (13%)	9.216	2.657 (22%)
14 Peach	1.216	0.024 (2%)	1.154	0.056 (5%)
23 Orange	75.892	41.960 (36%)	82.509	45.747 (36%)
General	6.253	1.943 (24%)	6.791	2.738 (29%)

Table 4.6: Results control models. Third iteration.

Variable	Description
hour	Time range. See Table 4.1.
weekday	Day of the week.
day	Day of the month. Kept only to order the dataset.
month	Month of the year.
year	Kept only to order the dataset.
section	Stops aggregation in each route.
route_name	Identifier of the different lanes in Ames' bus transport.
ons_yesterday	Total onboarding during the previous day for a certain section within a certain route in a certain hour. Only for onboarding prediction.
ons_last_week	Total onboarding during the previous week at the same weekday for a certain section within a certain route in a certain hour. Only for onboarding prediction.
offs_yesterday	Total offboarding during the previous day for a certain section within a certain route in a certain hour. Only for offboarding prediction.
offs_last_week	Total offboarding during the previous week at the same weekday for a certain section within a certain route in a certain hour. Only for offboarding prediction.
precip_mm	Inferred total rain and melted snow precipitation during the same hour range of the previous day.
snow	Inferred total snow precipitation during the same hour range of the previous day.
snow_d	Inferred inches of accumulated snow for the same hour range of the previous day.
ons	Target variable. Accumulated quantity of individuals that get on the bus at any stop of a certain section.
offs	Target variable. Accumulated quantity of individuals that get off the bus at any stop of a certain section.

Table 4.7: Variable description. Onboarding-Ogboarding dataset.

Variable	Description
hour	Time range. See Table 4.1.
weekday	Day of the week.
day	Day of the month. Kept only to order the dataset.
month	Month of the year.
section	Stops aggregation in each route.
route_name	Identifier of the different lanes in Ames' bus transport.
ons_yesterday	Total onboarding during the previous day for a certain section within a certain route in a certain hour.
ons_last_week	Total onboarding during the previous week at the same weekday for a certain section within a certain route in a certain hour.
precip_mm	Inferred total rain and melted snow precipitation during the same hour range of the previous day.
snow	Inferred total snow precipitation during the same hour range of the previous day.
snow_d	Inferred inches of accumulated snow for the same hour range of the previous day.
n_trips	Target variable. Total number of travels used to cover the demand at a specific time range for a route.
n_buses	Target variable. Total number of unique buses assigned to a route at a specific time range.

Table 4.8: Variable description. Resources dataset.

sources, the buses assigned to each route at a specific time range and the total number of trips done by all the assigned buses to cover the demand.

The pre-processing follows the one developed in Section 4.3. When aggregating by sections and time ranges, we made use of the trip and bus identifiers at each record to count the number of unique travels and vehicles, getting by that way both types of resources needed.

Due to the improvement of the processing developed in Section 4.3.1, it was decided to follow the same treatment, by adding columns for context trends, meteorological data, and the shifting average of the weather data.

With this, the final dataset of resources has a size of 11,644 records. A brief description of each field can be seen in Table 4.8.

4.4 Post-processing data analysis

Before conducting the model building, it is important to stop for a second to perform a brief data analysis of the process data, since we want to make sure that the steps taken correspond to the format we want to obtain.

As we have aggregated the stops into sections, we want to make sure that the strop grouping selection is correct, so we will create Figures similar to 4.3 and 4.4 to compare whether the distribution is similar or not. Figures 4.7 and 4.8 show that the stops grouping perfectly captures the essence of the stop-level data. For the first figure, we can clearly see the aforementioned popular onboarding zones; sections 2 and 3 represent the first group of stops at Lincoln Street, whereas section 4 captures the second popular zone at the second group of stops at Lincoln Street in the afternoon and evening. The same occurs for the second figure, where we can see the shift of morning-afternoon tendencies again; in the morning, the first sections are the most popular, while the inverse occurs as we pass the noon threshold.

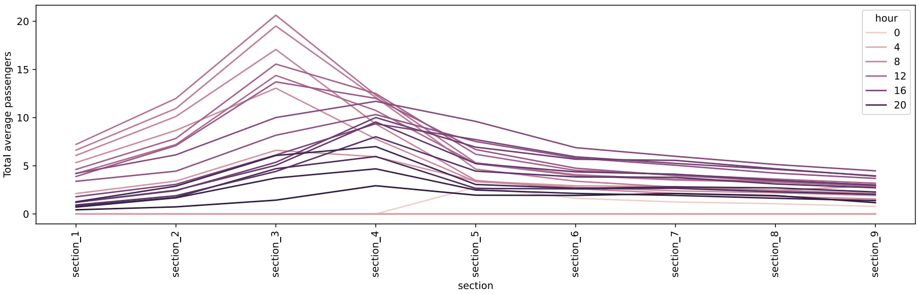


Figure 4.7: Average onboarding passengers per section for route 1 Red East

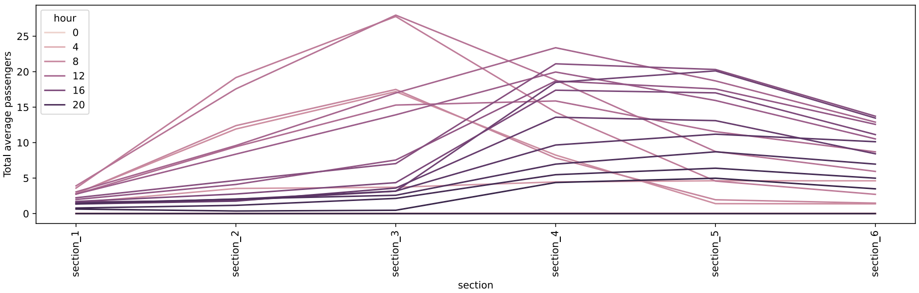


Figure 4.8: Average onboarding passengers per section for route 11 Cherry

Since we have created a couple of new variables for resource prediction, we want to know how they behave, since it may affect the scope of the prediction. Mainly, we want to know if the number of buses and trips assigned depends on the time period and section and, if so, how is that change. Figures 4.9 and 4.10 represent the evolution of the number of buses and trips, respectively, over the day for the different sections of route 1 Red West. The first thing we notice is that, as it was expected, the magnitude of the number of trips is larger than the number of buses, which has to be taken into account when performing a multioutput predictor model. The second thing we observe is that, as also expected, the assignation changes over time, which means that the bus system is somehow related to the passenger demand and the models we will create have to carry a temporal distinction of the time periods in Table 4.1. Lastly, we can recognize that the assignation quantity slightly changes depending on the section we observe; this makes much more sense if we relate this graphic to Figure 4.5 since those differences are clearly caused by bus transfers to other routes. We can even hypothesise about the larger gaps between sections as points where the transfers are taken place, but, as we do not dispose of this information, we can not know which points buses transfer in and which they transfer out. Therefore, are forced to ignore this behaviour and create resource prediction models for the whole route with time period distinction.

In this section, we have successfully transformed a raw data collection into multiple datasets enriched with external information, data depuration, and extracting internal relationships. Additionally, we have formatted the datasets in a way that we can create

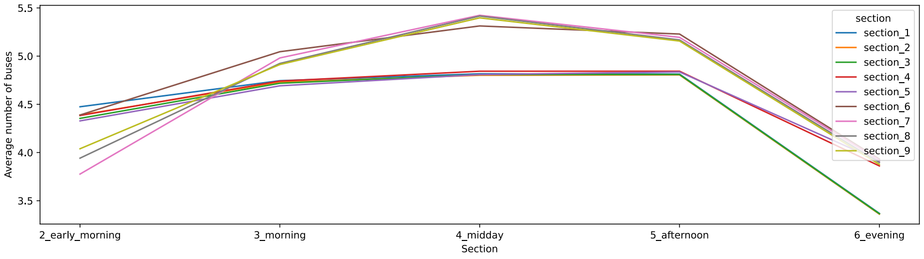


Figure 4.9: Average number of buses per section for route 1 Red West

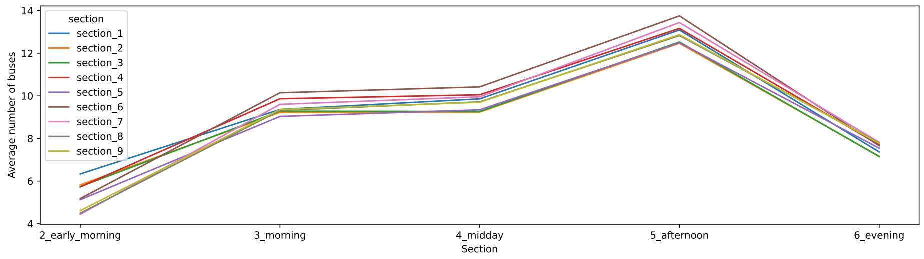


Figure 4.10: Average number of trips per section for route 1 Red West

complex models in a straightforward way. In the following section, we are going to see precisely this process of model creation.

CHAPTER 5

Knowledge extraction and model evaluation

In the following pages, we are going to talk about the experimentation process of the three tasks we want to predict: onboarding, offboarding, and resources. For all three, the different approaches or iterations realized to reach a final optimum model will be extensively explained, with its particular evaluation stage.

Additionally, we are going to discuss how these models are pipelined to produce traffic simulations.

5.1 Onboardings prediction

As our study uses a temporal segmentation of the data, we can consider two cases: whether the autocorrelation is kept or not, since we do not have the occupancy at each stop anymore, but an aggregation of the total occupancy within a period and sector. If the autocorrelation still exists in our data, treating it as a series problem will get appropriate results; we can do this since we still have a temporal order within the segmented data. On the other hand, if the autocorrelation is insufficient, we will opt for a more traditional approach treating each segment as an individual to be predicted via regression.

For the regression approach, we have blocks of individuals independent of each other, as they do not have the context of previous time steps. Nevertheless, in reality, the appearance of trends can happen, having a significant impact on the prediction capability of a given individual. The previous variables "ons_yesterday" and "ons_last_week" are used to guide the models giving the context of the normal and periodic trends, respectively.

We will consider the Random Forest (RF), Support Vector Machine for Regression (SVR) and Neural Network (NN) techniques for this approach. RF [5] combines multiple decision trees to create robust predictions, it has been chosen due to its capability to handle non-linear relationships and high-dimensional data, as in our case. SVR [7] creates a hyperplane that separates the individuals, as the RF, its capability to distinguish non-linear relationships and handle high-dimensional data has been of interest to us. Finally, the NN [4] learns the relationships among the variables to create its prediction, since it is highly customizable, we decided to iterate over multiple designs; the final one it's a dense NN with 6 hidden layers of sizes (64, 64, 32, 32, 16, 8) with a ReLU activation function in the layers and a linear one for the output layer.

Regardless of the technique, we will create an independent model for each route. This decision requires less time for model convergence and prediction than other solutions, such as, e.g., a general model with a route variable decision-maker.

In the series approach, the dataset is structured in independent series for each section within a route, which means that we will create a separate model for each section. In this case, we would not use the variables "ons_yesterday" and "ons_last_week", since the main feature of the models of this type is to use the own target variable as a look-back for predicting a new value, and we are not interested in have repetitive data, which translates into heavier and slower to train models. As in the previous methods, we want to capture the normal and periodic trends so that the look-back range will cover the previous week for a given individual. Apart from the target variable, we will make use of the variables "precip_mm", "snow", and "snow_d", since, as they are numeric data, they are also serial data that can be used as input for the models.

The LSTM and GM models will be used for series prediction. LSTM [21] is a type of RNN particularly effective for time-series regression tasks; similarly to the NN, it is highly customizable; we created a sequential model composed of an LSTM layer of size 100 and a dense output layer with a linear activation function. GM [13] is a widely used technique for passenger flow prediction since it focuses on finding the normal and periodic trends; in our case, we will use the variant GM(1,n), which uses other serial variables alongside the target one for predicting each new value.

5.1.1. Results

For the LSTM and GM models, the train-test split needs to be logical, which means that the test data must be the following time steps of the train data. This has been done in a proportion of an initial 80/20. After predicting a new time step, this value is introduced in the train split for predicting the following value.

On the other hand, the RF, SVR and NN models do not need that kind of split, but instead, it is preferable to perform a random split since we will cover a wider variability of the data by considering a high variety of time steps. Performing a re-training at each time stop would also be interesting and have similar results. Unfortunately, the temporary cost of it is enormous. The same 80/20 proportion will be considered for these models.

We want to compare how each model works for different routes to study their adaptability work regardless of the context of the route; nevertheless, each one has different magnitudes. The metric developed to assess that problem has been an R2 score calculated over the predicted values and the actual values of the test set. This metric calculates the similarity of two distributions. The range of this score covers from minus infinity to one; the closer to one, the better the model fits that route.

To compare models between them is preferable to use an absolute metric over the results, for which the mean absolute error (MAE) has been chosen, which calculates the absolute error between every pair of predicted values and the actual ones and performs the mean of all the errors.

The performance results are displayed in Table 5.1, where the green cells represent the best model for each route. We can see that the level of aggregation used has greatly eliminated the autocorrelation of the data, as the LSTM and GM not only are not the

Route	NN		RF		SVR		LSTM		GM	
	R2	MAE	R2	MAE	R2	MAE	R2	MAE	R2	MAE
1E	0.831	5.216	0.856	4.818	0.833	4.817	0.381	6.482	-0.024	7.259
1W	0.930	4.452	0.919	4.489	0.856	5.373	0.355	7.192	-1.453	14.766
2E	0.830	2.479	0.872	2.311	0.853	2.350	0.354	3.393	-0.180	3.564
2W	0.846	2.137	0.861	2.035	0.849	2.212	0.375	5.101	-0.665	7.078
3	0.927	12.112	0.912	12.196	0.902	14.098	0.543	14.984	-0.228	17.349
5	0.588	2.137	0.702	1.789	0.691	1.846	0.365	2.097	-0.194	2.137
7	0.777	3.544	0.820	3.268	0.813	3.666	0.445	4.035	-0.152	4.432
9	0.810	12.145	0.822	11.913	0.784	13.326	0.403	13.540	-0.316	54.321
11	0.940	7.371	0.935	6.867	0.913	9.037	0.542	12.593	-0.035	15.742
14	0.451	1.381	0.625	1.153	0.593	1.132	0.237	1.502	-0.073	1.680
23	0.881	83.662	0.892	72.366	0.884	20.201	0.837	81.472	-0.537	354.865

Wins	**3**	**2**	**8**	**7**	**0**	**2**	**0**	**0**	**0**	**0**

Table 5.1: Onboarding metrics performance. Routes: (1E: 1 Red East, 1W: 1 Red West, 2E: 2 Green East, 2W: 2 Green West, 3: 3 Blue, 5: 5 Yellow, 7: 7 Purple, 9: 9 Plum, 11: 11 Cherry, 14: 14 Peach, 23: 23 Orange)

best model for any route, but also the metrics for them are nowhere near the regression models; since LSTM averages a 43.97% for R2 score and 13.85 for MAE, and -35.06% and 43.93 respectively for GM.

Nevertheless, there is not a clear winner within the regression models since, depending on the criteria used, the best model for some routes differs. If both metrics agree on the model for a certain route, that should be the chosen model. That is the case of routes 1 Red West, 2 Green East, 2 Green West, 3 Blue, 5 Yellow, 7 Purple, 9 Plum and 23 Orange, which show values over 82% for the R2 score, except for 5 Yellow, which is 70.2%.

The rest of the routes show little difference in the metrics of the best-elected models, so the recommendation, in this case, is always to choose the model with a smaller MAE value. For routes 1 Red East and 14 Peach, the improvement in MAE is almost insignificant (0.001 and 0.021, respectively), but for 11 Cherry, the difference is 0.504.

According to the number of wins of each model, the RF technique results as the best model for both the R2 score and MAE, with 8 and 7 victories, respectively. In the second position comes the NN model, with 3 and 2, respectively. And finally, the SVR technique only results in the best model twice for the MAE metric.

From the results in Table 5.1, routes 5 Yellow and 14 Plum are notable for their low R2 score but also low MAE. The opposite happens with route 23 Orange, with an R2 score very close to one but a high MAE value.

Figures 5.1 and 5.2 show that the predictions fall within a small range. As the R2 score focuses on the similarity of the prediction distribution to the actual one, small errors within a narrow range translate into dissimilarity. Nevertheless, as our final focus is to assign buses according to the estimated passenger flow, any deviation within these ranges would not lead to significant resource shifts.

On the other hand, Figure 5.3 shows that the nature of this route is to have high variability; this distribution has been correctly captured according to the R2 score. Nevertheless, it still happens that any minor deviation in a prediction may cause major bus

Route	Election
1 Red East	SVR
1 Red West	NN
2 Green East	RF
2 Green West	RF
3 Blue	NN
5 Yellow	RF
7 Purple	RF
9 Plum	RF
11 Cherry	RF
14 Peach	SVR
23 Orange	RF

Table 5.2: Onboarding models assignation per route

rescheduling since, as seen in Table 5.1, the average error for the best case, RF model, is around 72, which is over a single bus capacity.

In Table 5.2, we can see the final model assignation of each route for the onboardings.

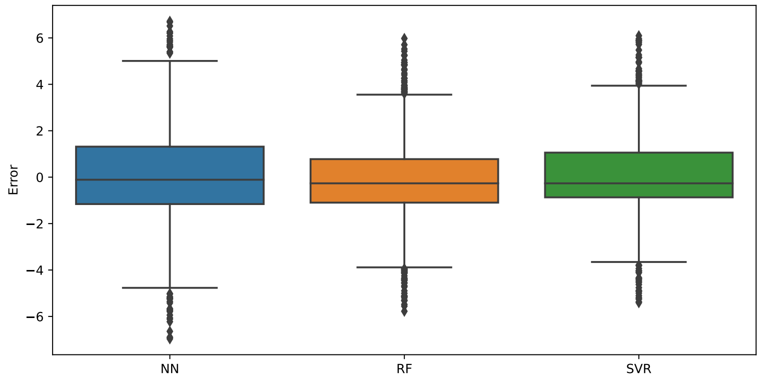


Figure 5.1: Boxplot model comparison for route 5 Yellow. Onboarding

5.2 Offboardings prediction

As we have the same data structure as in the previous chapter 5.1 and the behaviour of the target variable is similar, we will follow the same methodology with a minimal twist. As we have already seen in Table 5.1, the serial models underperform our problem due to the loss of self-correlation in the data. Therefore, we will not proceed with the analysis of those models.

5.2.1. Results

The criteria to train and test the data are the same as in Section 5.1.1. The results of the performance of these models are shown in Table 5.3, where the green cells represent the best model for each route. We can see the same phenomena as in Table 5.1, where, in some routes, different criteria select different models as the best ones. The routes that do

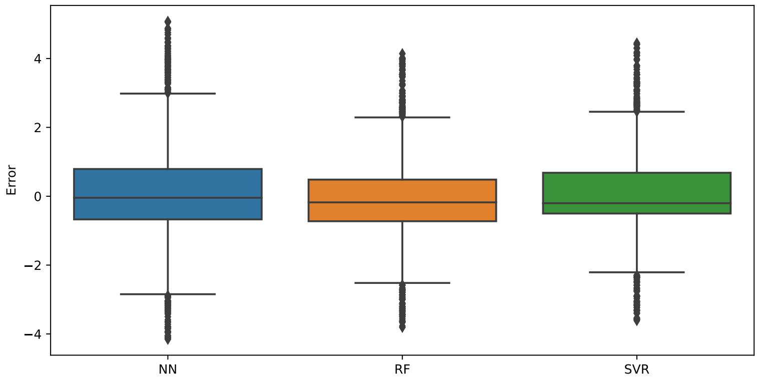


Figure 5.2: Boxplot model comparison for route 14 Peach. Onboarding

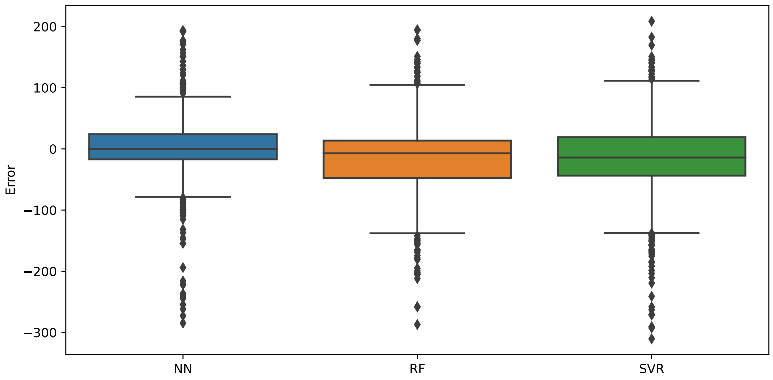


Figure 5.3: Boxplot model comparison for route 23 Orange. Onboarding

Route	NN		RF		SVR	
	R2	MAE	R2	MAE	R2	MAE
1E	0.859	4.311	0.871	4.061	0.823	4.750
1W	0.880	4.785	0.887	4.540	0.829	5.554
2E	0.866	2.404	0.871	2.296	0.813	2.659
2W	0.847	2.168	0.866	2.076	0.841	2.324
3	0.916	12.075	0.891	12.956	0.869	14.494
5	0.424	2.262	0.612	1.850	0.608	1.869
7	0.813	3.155	0.788	3.088	0.772	3.472
9	0.855	11.033	0.853	11.138	0.824	11.854
11	0.944	8.246	0.943	8.007	0.922	10.335
14	0.528	1.317	0.676	1.144	0.667	1.148
23	0.875	67.697	0.883	66.726	0.860	79.404

Wins	**4**	**2**	**7**	**9**	**0**	**0**

Table 5.3: Offboarding metrics performance. Routes: (1E: 1 Red East, 1W: 1 Red West, 2E: 2 Green East, 2W: 2 Green West, 3: 3 Blue, 5: 5 Yellow, 7: 7 Purple, 9: 9 Plum, 11: 11 Cherry, 14: 14 Peach, 23: 23 Orange)

not present this problem, in other words, that the metrics agree on the same model, are 1 Red East, 1 Red West, 2 Green East, 2 Green West, 3 Blue, 5 Yellow, 9 Plum, 14 Peach and 23 Orange, which show values over 85% for the R2 score, except for 5 Yellow and 14 Plum, which is 61.2% and 67.6%, respectively.

The routes that have more than one best-elected model, 7 Purple and 14 Peach, have almost an insignificant difference in the R2 metric, 2.5% and 0.1%, respectively, so it is preferable to be guided by the model with a lower MAE value, where the difference is of 0.067 and 0.239, respectively.

If we have a look at the number of wins of each model, the RF model is the clear winner for both metrics, with 7 victories for the R2 score and 9 for MAE. The NN technique comes in second place, resulting in the best model for four routes according to the R2 score and two for the MAE. The SVR model is not the best model in any of the cases.

As happened for the onboardings (see Chapter 5.1.1), in Table 5.3 we can see that routes 5 Yellow and 14 Plum have poor R2 scores, but the MAE error in both cases is low. Again, the opposite happens with route 23 Orange, where the R2 score shows pretty decent results, being over 0.88, while the MAE error highlights severe deviations over the prediction.

Same as in Figures 5.1 and 5.2, we can see in Figures 5.4 and 5.5 that, again, the range of the deviations is narrow enough to not cause significant changes in the resource assignation.

This similarity phenomenon also occurs in Figure 5.6, which can be related to its equivalent for onboardings (Figure 5.3), since we can see again a great variability in the deviations. Therefore, any minor prediction error may translate into severe bus rescheduling.

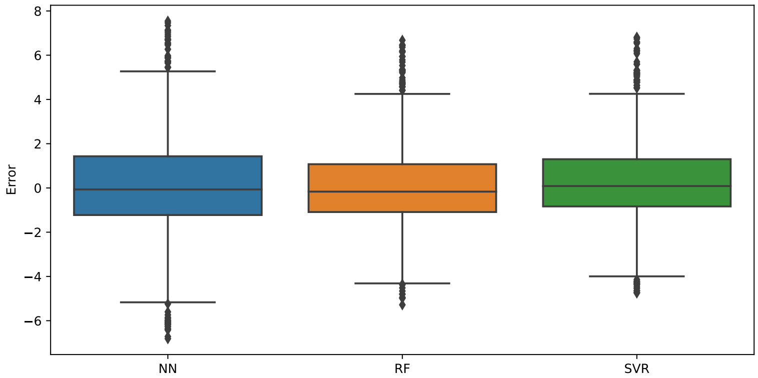


Figure 5.4: Boxplot model comparison for route 5 Yellow. Offboarding

Table 5.4 shows the final model assignation of each route for the offboardings.

5.3 Resources prediction

As seen in Table 4.8, there are two target variables, so we need to use techniques that let us get multiple outputs. It is necessary to get the predictions for both target variables from

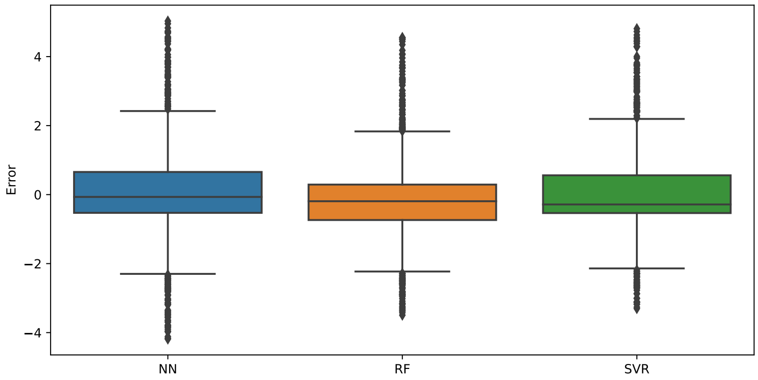


Figure 5.5: Boxplot model comparison for route 14 Peach. Offboarding

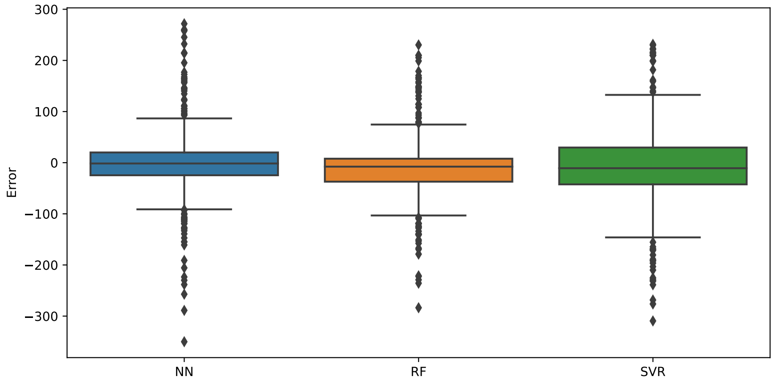


Figure 5.6: Boxplot model comparison for route 23 Orange. Offboarding

Route	**Election**
1 Red East	RF
1 Red West	RF
2 Green East	RF
2 Green West	RF
3 Blue	NN
5 Yellow	RF
7 Purple	RF
9 Plum	NN
11 Cherry	RF
14 Peach	RF
23 Orange	RF

Table 5.4: Offboardings models assignation per route

the same model since we want them to be related and avoid possible logical mistakes. A problem that can be induced from making separate predictions is to get a higher value for the target variable n_buses than for n_trips at the same timestamp, which is a waste of resources when applied to real life, as we will dedicate buses to a certain route that will never depart as the trips designed are filled. Therefore, we will stick with models and techniques which guarantee us having that relationship in a multi-output prediction. In order to arrive at a final model that obtains satisfactory results, a series of experiments and improvements have been carried out, as detailed in the following sections.

In order to compare how each model works for the different routes and for both target variables, we will use the accuracy if we are talking about a classification task and Mean Absolute Percentage Error (MAPE). This allows us to deal with the magnitude difference between routes and between the target variables by having a metric that can be compared on the same range, which is from 0 to 1. The reason why this metric has been chosen over the R2 score used in Chapter 5.1.1 is that R2 compares the distribution similarity of the predicted and actual data. In this case, as the dataset is smaller, the confidence of this metric over small datasets is not accurate enough to be used in this study.

To compare models between them is preferable to use an absolute metric over the results, for which the Root Mean Square Error (RMSE) has been chosen. Similar to the previous point, in Chapter 5.1.1 we used the MAE metric; here, it is avoided because of what an error in this prediction represents. If applied to real-life, an error of one unit over a bus or a trip equals to great losses of money, so we want to over-penalize these errors.

5.3.1. First version. Deciding the prediction task for each output

As seen in Figure 4.9, the range of values of the n_bus target variable is narrow when compared to the other target variable n_trips, Figure 4.10 This dimensionality difference can bias any multitarget model towards making predictions mainly based on the variable with a larger magnitude. This also inspired us to try different prediction tasks over the target variables to verify which is more precise.

Both strategies were based on using a NN model with a separated output layer for each target. The first one was inspired by the range of the target n_buses by treating it as a classification task, whereas for the target n_trips, as the range of values is greater, it was treated as a regression task. The second strategy was built in case the classification task did not perform well in predicting the number of buses, so instead, we dedicated it a regression layer. The structure of the NN is of 6 dense hidden layers of sizes (64, 64, 32, 32, 16, 16) with a ReLU activation function, one output layer branching from the second hidden layer of size 64 that makes the trips predictions, and a second output layer that is fed from the last hidden dense layer that serves for the buses predictions; these last two layers have a linear function as the activation function. Having the output layers at different levels of the NN is useful for making more precise predictions based on the dimensionality of each target variable.

As we can see in Table 5.5, there is no single doubt that treating both target variables as a regression task clearly gets better results than the alternative of a classification task for n_buses. This improvement is regardless of the route and target variable we analyze, so we will continue with this strategy on the following experimentation for improving the prediction task.

Route	Classification + Regression				Regression + Regression			
	n_buses		n_trips		n_buses		n_trips	
	Accuracy	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1 Red East	4.60%	2.053	0.168	1.273	0.135	1.483	0.123	0.799
1 Red West	2.31%	2.420	0.127	1.367	0.107	1.531	0.085	0.844
2 Green East	5.75%	1.655	0.130	1.223	0.102	0.878	0.061	0.469
2 Green West	4.21%	1.423	0.101	1.174	0.097	0.880	0.043	0.464
3 Blue	2.66%	2.615	0.162	2.250	0.203	1.871	0.075	0.680
5 Yellow	20.28%	0.964	0.073	0.531	0.114	0.847	0.042	0.316
7 Purple	27.15%	1.340	0.166	0.742	0.150	1.011	0.098	0.337
9 Plum	6.30%	2.663	0.134	0.967	0.186	1.834	0.075	0.451
11 Cherry	0.65%	3.061	0.138	1.925	0.226	1.927	0.052	0.701
14 Peach	19.02%	1.064	0.113	0.469	0.092	0.851	0.053	0.209
23 Orange	0.06%	12.674	0.165	3.359	0.461	8.199	0.076	1.903

Table 5.5: Results resources prediction version 1.

5.3.2. Second version. Technique experimentation

As seen in Table 5.5 the regression technique outperformed the classification one, so the following models will follow that philosophy. That is the case in this new set of experiments, where the multi-target regression strategy was tried. This technique fits a regressor per target variable. This ensures dedicating a particular predicting unit to each of the targets while keeping dependence between them so we do not relapse into the aforementioned problem of logical mismatches. The estimators that will be used with this strategy are the following:

Tikhonov Regularization The Tikhonov Regularization (TR) [22], also known as Ridge Regression, has a particular penalization term that reduces the probability of overfitting and overdependence of the target variables, which could be a great problem taking into account the magnitude difference of them, which will bias the model to make predictions mainly based on the variable with a larger magnitude, in this case, the number of trips.

RF Due to how well the RF technique worked with the onboardings and offboardings prediction (see chapters 5.1.1 and 5.2.1), it was considered a good idea to try this model, due to the nature similarity of the data.

SVR A similar idea was followed to build this model. Nevertheless, the SVR technique did not have results as good as RF. The reason why this model was contemplated is because of the ability of SVR to handle small datasets.

As shown in Table 5.6, the RF technique is the preferred one for both n_buses and n_trips in most of the routes. Nevertheless, in some cases, SVR and TR have similar or even better results for at least one of the target variables. To highlight the most extreme cases of that, we can talk about routes 2 Green West, where the difference between RF and SVR over the n_trips RMSE is 0.008, and 14 Peach, where the difference between both TR and SVR RMSE value with the one of RF at the target variable n_buses is of just 0.002. Furthermore, for route 5 Yellow, the SVR model surpasses the RF technique on the n_buses variable with a difference of 0.165; while for route 7 Purple, SVR is better than RF for both target variables, and even the TR technique outpaces RF on the n_buses variable.

Route		TR		RF		SVR	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
1 Red East	n_buses	0.950	0.194	0.714	0.077	0.826	0.103
	n_trips	1.851	0.227	1.331	0.107	1.478	0.132
1 Red West	n_buses	1.029	0.204	0.737	0.085	0.844	0.112
	n_trips	1.926	0.228	1.402	0.109	1.543	0.134
2 Green East	n_buses	0.619	0.141	0.406	0.050	0.479	0.056
	n_trips	1.288	0.200	0.841	0.088	0.904	0.095
2 Green West	n_buses	0.624	0.147	0.384	0.041	0.439	0.044
	n_trips	1.265	0.220	0.867	0.077	0.875	0.085
3 Blue	n_buses	0.814	0.288	0.601	0.111	0.623	0.128
	n_trips	2.323	0.295	1.707	0.126	1.772	0.166
5 Yellow	n_buses	0.634	0.079	0.478	0.061	0.313	0.405
	n_trips	0.946	0.191	0.829	0.100	0.887	0.123
7 Purple	n_buses	0.344	0.101	0.358	0.091	0.334	0.079
	n_trips	1.161	0.274	1.030	0.168	1.025	0.184
9 Plum	n_buses	0.402	0.102	0.346	0.057	0.373	0.081
	n_trips	1.517	0.192	1.181	0.123	1.329	0.151
11 Cherry	n_buses	0.819	0.188	0.579	0.088	0.689	0.102
	n_trips	2.266	0.174	1.490	0.098	2.866	0.123
14 Peach	n_buses	0.197	0.019	0.195	0.025	0.197	0.019
	n_trips	1.000	0.261	0.832	0.166	0.882	0.200
23 Orange	n_buses	1.636	0.357	1.022	0.137	1.379	0.266
	n_trips	8.351	0.351	4.788	0.150	6.678	0.253

Table 5.6: Results resources prediction version 2.

These slight differences make us think that not for all routes is there a clear winner, but the different approaches reach similar values using distant points of view. So in the next experimentation set, we will combine these techniques to create a definitive model.

5.3.3. Third version. Combining techniques

As seen in Table 5.6, some models perform better on predicting the number of buses than the number of trips, and vice-versa, and the same for the routes; so it was considered to take benefit from the best parts of each model by joining them into a single model. The first of them is not a model per se, but a simpler approach, where we averaged the predictions of each of the previous multioutput models weighted by the MAPE value of each one; we will refer to it as Avg. Ensemble. The second approach consists of a stacking ensemble, where the outputs of the previous regressors are used as the input of a new regressor to get a final prediction. The technique that will be used in this second-level prediction will be again the TR; this decision is to avoid again model biases over making a final prediction mainly based on the variable with a greater magnitude.

Table 5.7 shows the results of this last iteration. It is clear that averaging the Average Ensembler overall has better results than the Stacking one, as in most of the routes, both target variables show fewer errors for the beforementioned strategy. Nevertheless, in some routes, the best model for the target variable n_buses differ from the other one n_trips. That is the case of routes 5 Yellow, 7 Purple, 9 Plum, and 14 Peach.

Route		Avg. Ensembler		Stacking	
		RMSE	MAPE	RMSE	MAPE
1 Red East	n_buses	0.685	0.114	0.765	0.117
	n_trips	1.319	0.133	1.574	0.156
1 Red West	n_buses	0.724	0.122	0.854	0.137
	n_trips	1.389	0.134	1.458	0.138
2 Green East	n_buses	0.392	0.080	0.399	0.082
	n_trips	0.834	0.114	0.905	0.113
2 Green West	n_buses	0.370	0.067	0.377	0.085
	n_trips	0.795	0.102	0.888	0.129
3 Blue	n_buses	0.547	0.160	0.585	0.159
	n_trips	1.613	0.155	1.794	0.153
5 Yellow	n_buses	0.275	0.093	0.261	0.115
	n_trips	0.763	0.132	0.808	0.175
7 Purple	n_buses	0.287	0.129	0.341	0.122
	n_trips	0.915	0.196	0.825	0.178
9 Plum	n_buses	0.322	0.109	0.281	0.112
	n_trips	1.109	0.138	1.219	0.136
11 Cherry	n_buses	0.543	0.124	0.645	0.135
	n_trips	1.595	0.117	1.658	0.137
14 Peach	n_buses	0.176	0.040	0.199	0.046
	n_trips	0.783	0.207	0.739	0.210
23 Orange	n_buses	1.080	0.210	1.298	0.139
	n_trips	5.230	0.197	5.401	0.132

Table 5.7: Results resources prediction version 3.

For these routes, we have to decide which method we will keep with. It would be interesting to consider a solution based on minimizing the cost of over and under-allocation of both resources. Nevertheless, this would require further in-depth research, apart from the fact that we do not dispose of that kind of data. Therefore, we will use the Average Ensembler model regardless of the route selected, since it achieves better results overall and it requires less computation time than the alternative.

5.4 Simulation preparation

In order to explain how the models are pipelined and the data is formatted at each step, we have to explain first the tool we are going to use for the simulation process. This tool is called Simfleet [20], which is a simulation framework widely used for the evaluation and optimization of large-scale fleet management systems. The election of this simulation environment was a requirement since it was developed by VRAIN [1], the same research institute where this project has been conducted.

SimFleet is built on cutting-edge optimization techniques and integrates a variety of elements that have a significant impact on fleet operations, including traffic congestion, passenger demand, and vehicle routing. It is a versatile and expandable platform that enables the testing of multiple fleet management techniques, such as vehicle allocation, routing, and scheduling, in various scenarios and under varied limitations.

[1] https://vrain.upv.es/

SimFleet's capacity to produce realistic and dynamic scenarios while accounting for current traffic information and passenger demand is one of its key advantages. Because of this, it is especially helpful for assessing how well new technologies and fleet management algorithms function in actual, data-driven settings.

Based on a discrete event simulation model, SimFleet takes steps to update the system's state and start new events in response to certain occurrences, including the arrival of a passenger, the departure of a vehicle, or the occurrence of a traffic accident. A fixed time period (such as one minute) is represented by each discrete time step that the simulation takes.

The fleet management system's behaviour is simulated by SimFleet using a number of modules that communicate with one another. These modules include a network module that represents the transportation network (such as roads, intersections, and transit stops), a demand module that creates passenger requests and assigns them to vehicles, a vehicle module that simulates vehicle movement, and a dispatch module that assigns vehicles to passenger requests based on a set of rules or optimization criteria.

The user must first supply input data in the form of JSON files in order to use SimFleet to run a simulation. This data comprises information about the transport network's agents (such as the location, capacity, and travel time of each road segment), passenger demand (such as origin-destination pairings and desired arrival time), and fleet composition (such as the number and kind of vehicles).

The user may choose simulation settings such as the simulation's duration, the dispatch mechanism, and the performance metrics to be gathered once the input data has been entered into the simulation. Following then, the simulation proceeds in a series of time steps during which events are triggered, and measurements are made in accordance with the system's present condition.

SimFleet gathers information on a variety of system behaviours during the simulation, including vehicle utilisation, wait times for passengers, and system performance. Utilising built-in tools, this data may be evaluated and visualised, or exported for additional study.

Therefore, we have to define each of the individuals of the three types of agents we have in our system, bus stops, buses and passengers:

Bus stop The bus stops are static points in which passengers and buses will interact, so they have no other information rather than the physical location, in other words, the latitude and longitude of each one. This information is absent from the original data 4.1, so we have to gather it from other sources. For this purpose, we have used the Google Maps API [2]. This API lets us retrieve multiple data fields from Points of Interest (POI), which in our case will be the bus stations, from which we are only interested in the geographical data. Each API request consists of a query representing the POI to obtain, and the output is a list of items that match the query. To refine the data-gathering process, each request consisted of the name of the bus stop plus 'bus stop', since the name of some bus stops may match other POI, and this will help to restrict the outcome entities to the category we are interested in; additionally, we included an optional location bias field to restrict the results to the area of the city of Ames. Out of the 356 bus stops in our bus system, only for 3 of them, this information could not be automatically gathered, for which, using

[2] https://developers.google.com/maps?hl=en

the live online version of the bus system map [3], the approximate coordinates were extracted manually from Google Maps as accurate as possible.

Bus The buses are mobile agents, which means that their location and state change over time. Nevertheless, this is a behaviour internally implemented on SimFleet, so we only have to worry about the initial state of each agent.

First of all, we have to determine how many buses will be assigned to a certain setup based on the system characteristics. For this, we use the final models developed in Chapter 5.3.3. We will then get the number of buses that have to be assigned to each route at each time period, as well as the number of trips that the original bus system will design for the experiment setup conditions. It is worth noting that the trips will are not included in the simulation per se; instead, they are used to be compared to those needed in the simulation to analyze possible resource waste in the bus system.

The starting point of the predicted buses will be uniformly distributed along the route, since we do not have the starting point of each route in the data, especially on the circular routes, and due to SimFleet design restrictions, it is complex to spawn a bus at the same starting point at a given time delay each. Additionally, all buses will have a total capacity of 60 passengers, as it is the most common bus type, as seen in Chapter 4.1.

Passenger Same as the buses, the passengers are mobile agents, where the only attribute that is changed during the simulation is their location, which will be based on the bus the passenger is on. Again, this behaviour is controlled by SimFleet, and we only have to describe the onboarding and offboarding stops of each passenger. It is worth mentioning that, for practical uses, we will consider that each passenger takes the service only once, regardless of the time span of the experimental simulation.

The first step is to determine the number of passengers per section. This is the reason why we built the onboarding predictions model in Chapter 5.1. Using the best model for each route as highlighted in Table 5.2, we will get the number of passengers that will get on the bus for a specified route, section of the route and time period. Both sections and time ranges are levels of aggregation that were needed to build proper data models, but now it is an inconvenience for the simulation since we want to give it more granularity, so we have to distribute the load in an appropriate manner.

With the original data, within each route, for each section and time period, we calculate the cross-distribution proportion of each stop and 30-minute period within the selected section and time period, using all the historical data of onboardings. The same has been done for the offboardings.

Using this distribution data, we disaggregate each section and time period prediction to each of its stop and 30-minute period sub-levels by multiplying the amount predicted per the respective distribution value. With this, in the simulation, we will be injecting passengers at each stop in a phased manner, achieving then a more accurate representation of a real-world situation.

We have just obtained the initial position and delay of each passenger, but we yet need to know when each one will get off the bus. Unfortunately, in the data, it is not tracked the journey that each passenger takes; in other words, where he/she

[3] https://www.mycyride.com/map

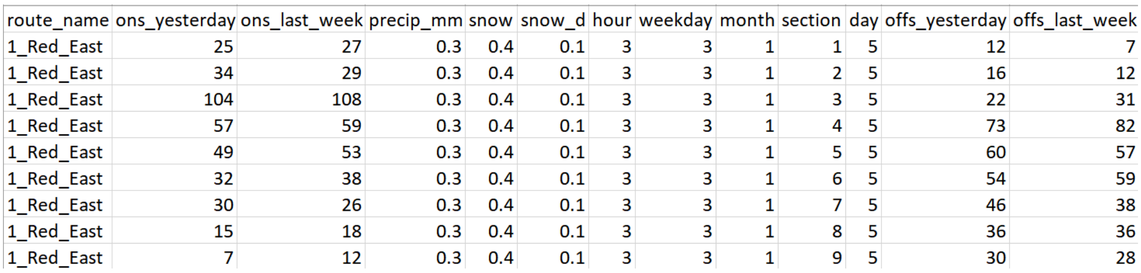
gets on and off the bus; instead, we have an aggregate value. For this reason, using the Large Numbers Law [8], we will assume the essence of the passengers' mobility is captured overall.

Here, we can not directly use the prediction of offboardings of each model, since there may be slight mismatches with the amounts predicted for the onboardings. Instead, we will use a probability to assign each passenger's destination stop. For this process, firstly, we have to predict the number of passengers that the best-elected models for each route (see Table 5.4) assign to all the sections of a route given a time period since passengers can take off the bus at any stop of the route. It is worth mentioning that, in the linear routes, passengers can only get off the bus at stops subsequent to the one they got on; also, we will consider all of the offboarding at the same time period as the onboardings. Then, we redefine the prediction of each section and time period to a proportion of the total predicted for the section, since we want to capture the particular distribution of the experimental time span, and we disaggregate that value by the proportions beforementioned at the historical offboarding distribution data. With this, we have the probability of a passenger getting off the bus at each bus stop at a given 30-minute period, so, using these probabilities as the weights of a random selector, we assign each passenger the destination stop.

All these processes are already pipelined and prepared to format the data at each point correctly to produce the desired outcome. The input for this procedure is a CSV file formatted as described in 4.7, and the output we get is a JSON file with the format and fields that SimFleet expects as the load data.

5.4.1. Simulation example

As a final statement, we want to show an example of an experiment correctly loaded on SimFleet. For practical uses, we will limit the scope of the experiment to a particular route, day and time period.

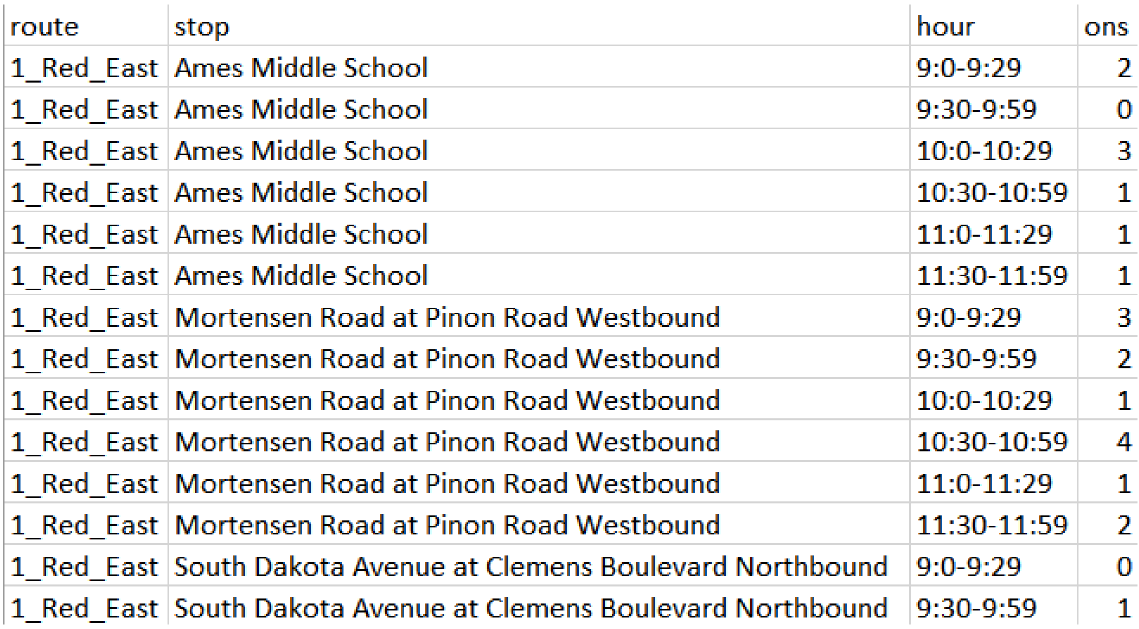


route_name	ons_yesterday	ons_last_week	precip_mm	snow	snow_d	hour	weekday	month	section	day	offs_yesterday	offs_last_week
1_Red_East	25	27	0.3	0.4	0.1	3	3	1	1	5	12	7
1_Red_East	34	29	0.3	0.4	0.1	3	3	1	2	5	16	12
1_Red_East	104	108	0.3	0.4	0.1	3	3	1	3	5	22	31
1_Red_East	57	59	0.3	0.4	0.1	3	3	1	4	5	73	82
1_Red_East	49	53	0.3	0.4	0.1	3	3	1	5	5	60	57
1_Red_East	32	38	0.3	0.4	0.1	3	3	1	6	5	54	59
1_Red_East	30	26	0.3	0.4	0.1	3	3	1	7	5	46	38
1_Red_East	15	18	0.3	0.4	0.1	3	3	1	8	5	36	36
1_Red_East	7	12	0.3	0.4	0.1	3	3	1	9	5	30	28

Figure 5.7: Input example.

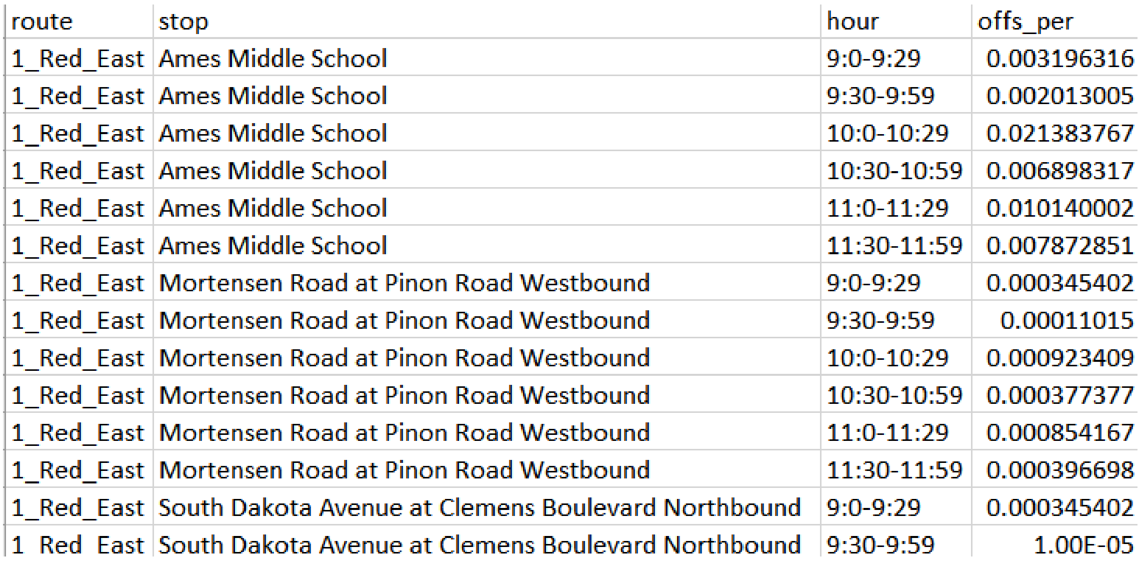
Figure 5.7 shows the experimental conditions we are using in the simulation load. In this case, we have decided to show route 1 Red East for being the largest route in terms of sections, and, because of being linear, the visualization will be more comfortable. The columns 'ons_yesterday' and 'ons_last_week' have been filled following the passenger distribution in Figure 4.7, same as for 'offs_yesterday' and 'offs_last_week'. The environmental columns ('precip_mm', 'snow' and 'snow_d') have been filled with a random value each since we are not interested in experimenting with the effect of those values in the traffic, but just making sure the simulation data load works; it is worth noting that, as mentioned in Chapter 4.3.1, we only use one meteorological station for the whole data, so we can not differentiate the environmental conditions at different physical places of the route. As for the time variables ('hour', 'weekday', 'month' and 'day'), the values have also been chosen randomly for the same reason; we are not interested in giving a significant value for experimentation, rather than having any value for the data load; here it is worth noting that the variable 'day', although not being used in the models, it is needed to keep the input data sorted. Lastly, in the column 'section', we included all the sections, one individual for each.

This input data is formatted to create the two datasets that will be used by the models, one for the onboarding and offboarding models, and another for the resources prediction. These datasets are directly pipelined to the best-elected models for each of the tasks and routes, which create the predictions. These predictions are afterwards formatted to disaggregate the sections into stops and the time periods into 30-minute periods.



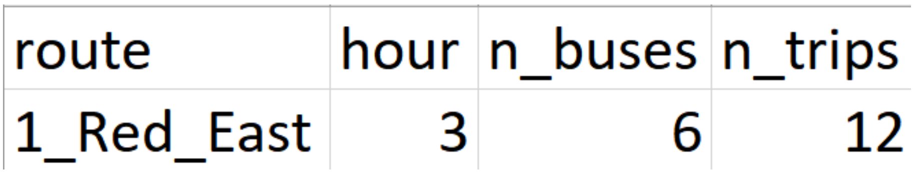
route	stop	hour	ons
1_Red_East	Ames Middle School	9:0-9:29	2
1_Red_East	Ames Middle School	9:30-9:59	0
1_Red_East	Ames Middle School	10:0-10:29	3
1_Red_East	Ames Middle School	10:30-10:59	1
1_Red_East	Ames Middle School	11:0-11:29	1
1_Red_East	Ames Middle School	11:30-11:59	1
1_Red_East	Mortensen Road at Pinon Road Westbound	9:0-9:29	3
1_Red_East	Mortensen Road at Pinon Road Westbound	9:30-9:59	2
1_Red_East	Mortensen Road at Pinon Road Westbound	10:0-10:29	1
1_Red_East	Mortensen Road at Pinon Road Westbound	10:30-10:59	4
1_Red_East	Mortensen Road at Pinon Road Westbound	11:0-11:29	1
1_Red_East	Mortensen Road at Pinon Road Westbound	11:30-11:59	2
1_Red_East	South Dakota Avenue at Clemens Boulevard Northbound	9:0-9:29	0
1_Red_East	South Dakota Avenue at Clemens Boulevard Northbound	9:30-9:59	1

Figure 5.8: Onboarding predictions formatted



route	stop	hour	offs_per
1_Red_East	Ames Middle School	9:0-9:29	0.003196316
1_Red_East	Ames Middle School	9:30-9:59	0.002013005
1_Red_East	Ames Middle School	10:0-10:29	0.021383767
1_Red_East	Ames Middle School	10:30-10:59	0.006898317
1_Red_East	Ames Middle School	11:0-11:29	0.010140002
1_Red_East	Ames Middle School	11:30-11:59	0.007872851
1_Red_East	Mortensen Road at Pinon Road Westbound	9:0-9:29	0.000345402
1_Red_East	Mortensen Road at Pinon Road Westbound	9:30-9:59	0.00011015
1_Red_East	Mortensen Road at Pinon Road Westbound	10:0-10:29	0.000923409
1_Red_East	Mortensen Road at Pinon Road Westbound	10:30-10:59	0.000377377
1_Red_East	Mortensen Road at Pinon Road Westbound	11:0-11:29	0.000854167
1_Red_East	Mortensen Road at Pinon Road Westbound	11:30-11:59	0.000396698
1_Red_East	South Dakota Avenue at Clemens Boulevard Northbound	9:0-9:29	0.000345402
1 Red East	South Dakota Avenue at Clemens Boulevard Northbound	9:30-9:59	1.00E-05

Figure 5.9: Offboarding predictions formatted



route	hour	n_buses	n_trips
1_Red_East	3	6	12

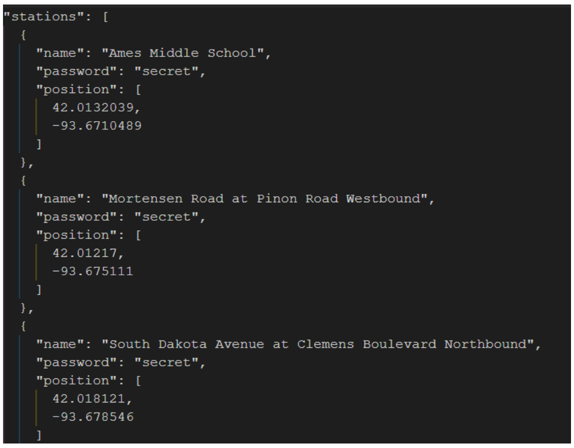
Figure 5.10: Resources predictions formatted

Figure 5.8 shows the data formatted after the predictions. As we can see, we have a prediction for each stop and 30-minute period level. This will dictate the number of passengers that will spawn at each stop at each 30-minute period.

In Figure 5.9, we have the same level of aggregation as for the onboardings, with the exception that the predicted value is not the number of passengers that take off the bus at each stop and 30-minute period, but the probability of the stop to being elected as the destination one at a given time period. It is worth noting that the 30-minute periods are independent of each other, so when electing the destination stop of a passenger, we will only take the ones of the same period as when the passenger took the bus.

Figure 5.10 shows the data formatted for the resources. In this case, we only have one record, as we only took one route and one time period for the experimentation. The buses predicted are supposed to be on duty during the entire experimentation, regardless of the number of trips they need to supply the demand, as the variable ’n_trips’ is only used to be compared with its simulated alike.

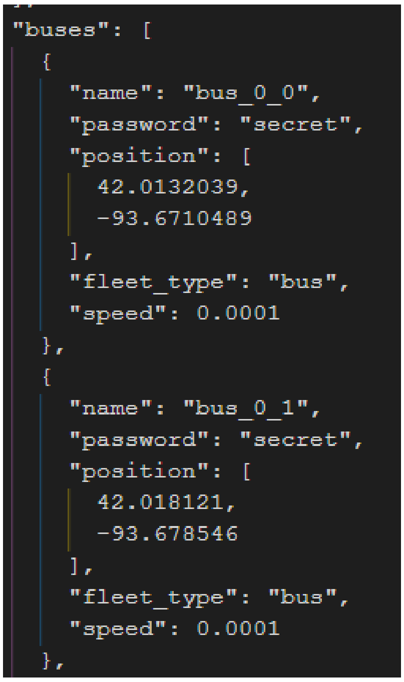
Using these formatted datasets, we have to transform them with the format of the SimFleet input JSON file, following the steps described in Chapter 5.4. With this, we will obtain the individual bus stations, the buses and the passengers in the simulation. The final JSON is showed divided into each of the three types of agents in Figures 5.11, 5.12 and 5.13.



```
"stations": [
  {
    "name": "Ames Middle School",
    "password": "secret",
    "position": [
      42.0132039,
      -93.6710489
    ]
  },
  {
    "name": "Mortensen Road at Pinon Road Westbound",
    "password": "secret",
    "position": [
      42.01217,
      -93.675111
    ]
  },
  {
    "name": "South Dakota Avenue at Clemens Boulevard Northbound",
    "password": "secret",
    "position": [
      42.018121,
      -93.678546
    ]
```

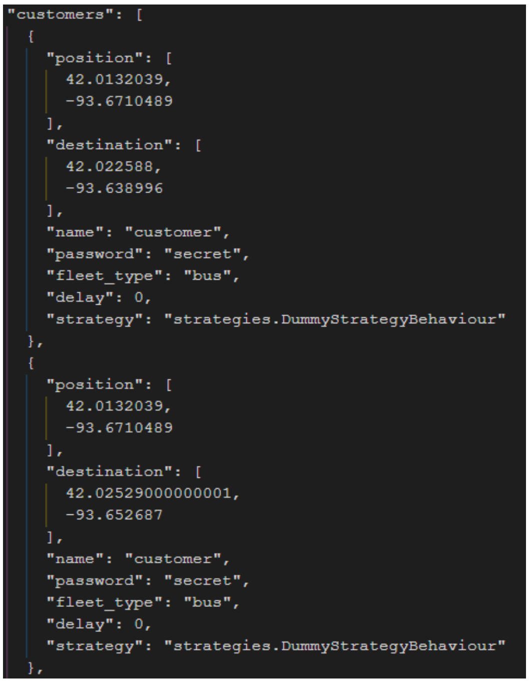
Figure 5.11: Bus stations JSON load

In Figure 5.14, we can see the initial state of the simulation, where all the agents (stations, buses and passengers) are loaded in the simulation as dictated by the predictions. Unfortunately, we can not show how the different agents interact in the simulation via passengers getting on and off the bus and the vehicles moving across the route. This is due to the behaviour of the bus not yet being implemented by SimFleet’s developers by the time of the writing of this memory; we expect it to be built any time soon to continue the project experimenting with the system’s behaviour under different circumstances to detect its vulnerabilities.



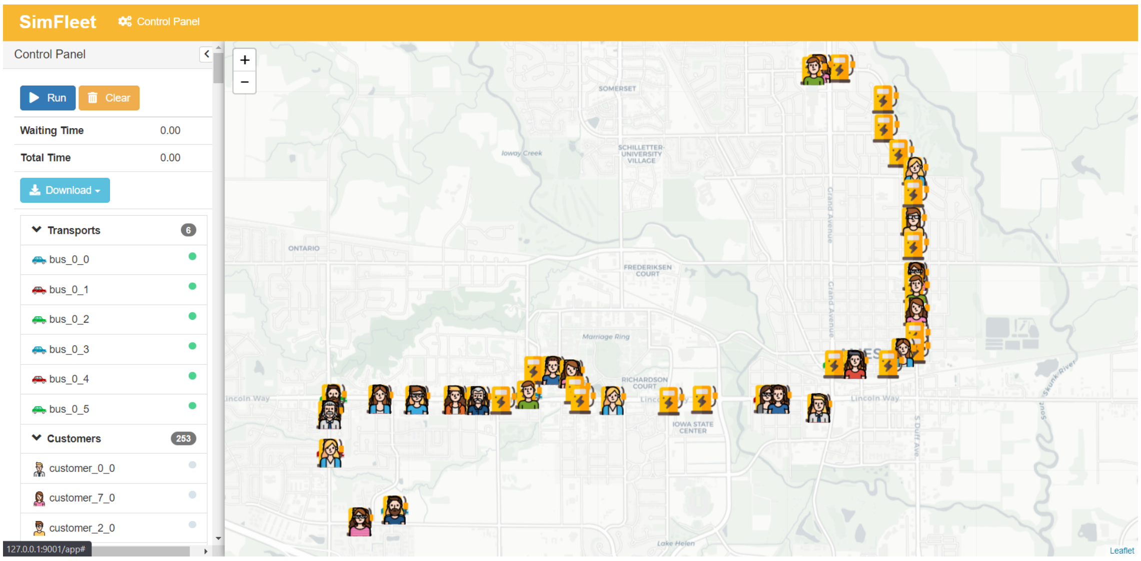
```
"buses": [
  {
    "name": "bus_0_0",
    "password": "secret",
    "position": [
      42.0132039,
      -93.6710489
    ],
    "fleet_type": "bus",
    "speed": 0.0001
  },
  {
    "name": "bus_0_1",
    "password": "secret",
    "position": [
      42.018121,
      -93.678546
    ],
    "fleet_type": "bus",
    "speed": 0.0001
  },
```

Figure 5.12: Buses JSON load



```
"customers": [
  {
    "position": [
      42.0132039,
      -93.6710489
    ],
    "destination": [
      42.022588,
      -93.638996
    ],
    "name": "customer",
    "password": "secret",
    "fleet_type": "bus",
    "delay": 0,
    "strategy": "strategies.DummyStrategyBehaviour"
  },
  {
    "position": [
      42.0132039,
      -93.6710489
    ],
    "destination": [
      42.02529000000001,
      -93.652687
    ],
    "name": "customer",
    "password": "secret",
    "fleet_type": "bus",
    "delay": 0,
    "strategy": "strategies.DummyStrategyBehaviour"
  },
```

Figure 5.13: Passengers JSON load



Waiting Time 0.00
Total Time 0.00

Transports 6
- bus_0_0
- bus_0_1
- bus_0_2
- bus_0_3
- bus_0_4
- bus_0_5

Customers 253
- customer_0_0
- customer_7_0
- customer_2_0

Figure 5.14: Simulation initial state

CHAPTER 6
Conclusions

In conclusion, this work emphasizes the importance of effective management of public transportation systems to provide passengers with high-quality service. One key factor in achieving this goal is the accurate prediction of bus occupancy levels, which can help optimize bus routes, increase service reliability, and decrease passenger wait times. However, there is a limited amount of research in this area, with most existing studies focusing on either real-time sensor data or historical bus occupancy data. Therefore, this study aimed to set a benchmark for bus and other public transport systems by using both traditional and machine learning techniques to develop models that can predict bus occupancy levels with high accuracy. The main contribution of the study is to propose a methodology for predicting passenger demand and simulating bus system traffic to characterize the vulnerabilities of the service.

The goals achieved during this project are as follows:

- Firstly, the proposed solution was carefully chosen by identifying the most suitable dataset that provides the necessary information for our analysis and modelling via an extensive analysis of publicly available datasets on public urban transportation, extracting valuable insights used in our methodology in a research study of previous works related to urban transport passenger demand, and analyzing different simulation techniques to select the most appropriate one based on our requirements.

- Related to the previous point, by conducting a comprehensive study of the current state of public transport, we could identify the weaknesses and areas that hold potential for improvement, guiding our efforts towards addressing the misuse of resources to enhance the performance of bus transport systems.

- An in-depth dataset characterization allowed us to identify the key aspects that impacted the subsequent processing stage, where the data was given an appropriate format for each of the modelling purposes. The three tasks outlined, which are passenger onboardings and offboardings; and resource allocation, were successfully approached via carefully designed models to achieve accurate and reliable predictions.

- We found that aggregating the data led to making it unsuitable for treating the problems as a series. This is because, by creating time segments during the preprocessing, the capability of self-explainability in the series has been lost. Nevertheless, we cannot assume that regardless of the level of aggregation used, the phenomenon remains the same, although, the study of the effect on smaller aggregation ranges is outside our scope since it is not useful for our purpose.

- Additionally, since we want the predictions to serve as input of a simulation environment, it is more beneficial for our purpose to use regression instead of serial techniques, since the first methods let us, with the given data, create any environment with few experimental data, while it is needed to continue the series up to the experimental point to recreate the conditions wanted in the second case.

- Based on the data achieved from the bus network of Ames, Iowa, it has been observed that, among the proposed, there is not a regression model that is above the rest; since, depending on the particularities of each route, each requires a customized study to optimize the system. Therefore, if applied to other routes in this network or others, it will be essential not to follow the models developed in this project blindly and try different methodologies.

- Lastly, to facilitate the integration of predictions into a simulation, we have created a pipeline that formats the predictions into an interpretable format suitable for simulation purposes. This pipeline ensures full compatibility between the prediction models and the simulation framework.

It is worth noting that part of this project has been accepted for being published in a conference article. The segment in question deals with the handling, processing and modelling of the upstream prediction. The paper will be found as follows:

A. Ibáñez. J. Jordán, V. Julian, Improving Public Transportation Efficiency through Accurate Bus Passenger Demand. Workshop on Adaptive Smart areaS and Intelligent Agents at PAAMS 2023. Guimarães, 12-14 July 2023.

In conclusion, this project has successfully accomplished a range of objectives that serve as a step forward in the understanding of public urban transportation systems and developing data-driven solutions that can contribute to their sustainable and efficient operation.

6.1 Future work

While this project has achieved significant milestones in analyzing urban transport services through demand forecasting, there are several avenues for future exploration and improvement. The following are potential areas for further research and development:

Enhanced data collection Future work should focus on expanding the scope of data collection efforts. This could involve obtaining additional datasets that capture a broader range of factors influencing passenger demand, such as special events, or socioeconomic indicators, among others. By incorporating more comprehensive data sources, the accuracy and robustness of the predictive models can be further improved.

Advanced modelling techniques While this project has employed state-of-the-art modelling techniques, there is room for exploring more advanced algorithms and methodologies. Future research can investigate the application of more advanced machine learning techniques, such as deep learning or ensemble models, to further improve the accuracy and precision of the demand forecasting models.

Policy and implementation guidelines As the project outcomes can have significant policy implications, future work should involve the development of guidelines and recommendations for policymakers and transportation authorities. These guidelines can help in the effective implementation of data-driven strategies, highlighting the potential benefits and addressing potential challenges associated with their adoption.

Integration of demand-responsive solutions The development of demand-responsive transportation solutions presents an exciting avenue for future work. Our project is based on static data, which in the long term may cause the models to be outdated and, therefore, the decisions taken based on that not to respond to the up-to-date passenger demand. By incorporating dynamic routing and scheduling algorithms into the models, it becomes possible to optimize resource allocation based on real-time demand patterns. This would enable transportation providers to adapt their services in response to fluctuating passenger demand, resulting in improved efficiency and customer satisfaction.

Scalability and transferability To ensure the practical implementation of the developed models and methodologies, future research should focus on scalability and transferability. This would involve exploring methods for scaling up the models to larger transportation networks and evaluating their performance in different geographical contexts. Additionally, consideration should be given to the adaptability and transferability of the developed solutions to diverse urban environments and transportation systems.

6.2 Relation of the work developed with the studies pursued

This project owes its success to the data science degree pursued at this university. The comprehensive education received and the skills honed over the course of four academic years have empowered me to apply and derive value from this knowledge in practice effectively. The following points will highlight the subjects that played a key role in the success of each of the processes:

- The analysis, cleaning and data preprocessing have been skills strongly developed in subjects such as Exploratory Data Analysis, and Projects I to III.

- All the models developed have been the outgrowth of a wide range of methods and abilities learned from subjects such as Statistical Models for Decision-making I and II, Descriptive and Predictive Models I and II, Projects I to III, and Evaluation, Deployment and Monitorisation of Models.

- The general programming practices and pipelining methods were effectively employed thanks to Programming and Algorithmics.

- The appealing visuals provided throughout the whole memory have been a technique highly improved thanks to Visualization.

- Finally, the API used in Section 5.4 meant a huge time saving from gathering the stops coordinates that would have been manually otherwise. This is the result of a combination of expertise I could learn in Programing and Data Acquisition and Transmission.

Bibliography

[1] Seoul information on the number of people getting on and off at each bus route by bus route and by time zone (including the location of the stop). Korean Public Data Portal, 2022. Data retrieved from: https://www.data.go.kr/en/data/15100899/fileData.do.

[2] A. Baghbani, N. Bouguila, and Z. Patterson. Short-term passenger flow prediction using a bus network graph convolutional long short-term memory neural network model. pages 1331–1340, 2023.

[3] J.P. Barreto, H. Novak, and R. Souza. São paulo bus system. 2020. Data retrieved from: https://www.kaggle.com/datasets/joaofb/so-paulo-bus-system?select=overview.csv.

[4] C.M. Bishop. Neural networks for pattern recognition. In *Oxford university press*, 1995.

[5] L. Breiman. Random forests. volume Vol 45, pages 5–32, 2001.

[6] Toronto Transit Commission. Real-time bus occupancy information. Information retrieved from: https://www.ttc.ca/riding-the-ttc/Real-Time-Bus-Occupancy-Info.

[7] C. Cortes and V. Vapnik. Support-vector networks. volume 20, pages 273–297, 1995.

[8] P. Erdös and A. Rényi. On a new law of large numbers. volume 22, pages 103–111, 1970.

[9] Transport for London. Transport for london apis. Infromation retrieved from: https://api-portal.tfl.gov.uk/api-details#api=Line&operation=Forward_Proxy.

[10] Transport for NSW. Bus occupancy. 2017. Data retrieved from: https://opendata.transport.nsw.gov.au/dataset/bus-occupancy-aug-2016-jan-2017.

[11] F. Gunawan, S. Suharjito, and A. Gunawan. Simulation model of bus rapid transit. volume 68, 2014.

[12] B. Hajinasab, P. Davidsson, J. Persson, and J. Holmgren. Towards an agent-based model of passenger transportation. pages 132–145, 2016.

[13] D. Julong. Introduction to grey system theory. volume 1, pages 1–24, 1997.

[14] Yang Liu, Zhiyuan Liu, and Ruo Jia. Deeppf: A deep learning based architecture for metro passenger flow prediction. volume 101, pages 18–34, 2019.

[15] S. Liyanage, R. Abduljabbar, H. Dia, and P. Tsai. Ai-based neural network models for bus passenger demand forecasting using smart card data. volume 11, pages 365–380, 2022.

[16] T.K. Madsen, H.C. Schwefel, and L.M. Mikkelsen. Live deployment data of bus occupancy. VBN, 2022. Data retrieved from: doi.org/10.5278/48b627f9-f45b-4cbd-a085-9046fb1425fe.

[17] W. Ming, Y. Bao, Z. Hu, and T. Xiong. Multistep-ahead air passengers traffic prediction with hybrid arima-svms models. volume 2014, 2014.

[18] T. Moyo, A. Kibangou, and W. Musakwa. Bus network and occupation survey in johannesburg. IEEE Dataport, 2021. Data retrieved from: https://dx.doi.org/10.21227/xnv8-zh36.

[19] N. Nagaraj, H.L. Gururaj, B.H. Swathi, and Y. Hu. Passenger flow prediction in bus transportation system using deep learning. volume 81, pages 12519–12542, 2022.

[20] J. Palanca, A. Terrasa, C. Carrascosa, and V. Julián. Simfleet: A new transport fleet simulator based on mas. In *Highlights of Practical Applications of Survivable Agents and Multi-Agent Systems. The PAAMS Collection*, pages 257–264. Springer International Publishing, 2019.

[21] J. Schmidhuber and S. Hochreiter. Long short-term memory. volume 9, pages 1735–1780, 1997.

[22] N. Tikhonov and V.Y. Arsenin. Solutions of ill-posed problems. In *Wiley*, 1977.

[23] TransLoc. Transloc publicapi. 2021. Infromation retrieved from: https://rapidapi.com/transloc/api/openapi-1-2/details.

[24] Xiaoyuan Wang, Yongqing Guo, Chenglin Bai, Shanliang Liu, Shijie Liu, and Junyan Han. The effects of weather on passenger flow of urban rail transit. volume Vol 6, No 1, pages 11–20, 2020.

[25] K. Wilbur. CyRide Automatic Passenger Counter Data, October 2021 - June 2022. 2022.

[26] Z. Zhang, X. Xu, and Z. Wang. Application of grey prediction model to short-time passenger flow forecast. volume 1839, 2017.

APPENDIX A
Project relationship with the Sustainable Development Goals

This project can be related to the United Nations' Sustainable Development Goals (SDGs). In Table A.1, we can see how this project is related to each of the goals with its degree of adequation. The strongest bondings are with goals 9 (Industry, innovation, and infrastructure), 11 (Sustainable cities and communities), and 13 (Climate Action). The reasoning is as follows:

Goal 9 This project aligns with Goal 9 by promoting innovation and the development of sustainable infrastructure within the urban transport sector. By employing data-based strategies and advanced modeling techniques, the project aims to enhance the efficiency and effectiveness of urban transport systems, contributing to the overall improvement of transportation infrastructure. While the project directly addresses the innovative use of data and technology, its impact on broader industrial and infrastructural aspects may vary depending on the specific implementation and scale of the project.

Goal 11 The project strongly relates to Goal 11, which focuses on creating sustainable cities and communities. By analyzing urban transport systems and developing data-driven solutions, the project aims to address challenges related to congestion, pollution, and resource mismanagement. Through improved resource allocation, operational efficiency, and congestion mitigation, the project contributes to the creation of sustainable, efficient, and livable urban environments. The outcomes of the project have the potential to positively impact the quality of life for urban residents, enhance accessibility, and promote environmentally-friendly transportation options.

Goal 13 While the project indirectly relates to Goal 13, its impact on climate action is significant. By optimizing resource allocation, reducing congestion, and promoting the use of public transportation, the project contributes to reducing greenhouse gas emissions associated with urban transportation. The enhanced efficiency and reliability of public transport systems encourage modal shifts from private vehicles to buses, leading to decreased reliance on fossil fuels and lower carbon emissions. The project's impact on climate action may vary depending on the specific context, scale, and implementation of the data-driven solutions within the transportation sector.

It is important to note that the level of relevance to the SDGs may vary depending on the specific objectives, implementation, and scale of the project. While the project

Sustainable Development Goal	**High**	**Medium**	**Low**	**N.A.**
1. End of poverty.				X
2. Zero hunger.				X
3. Health and wellbeing.				X
4. Quality education.				X
5. Gender equality.				X
6. Clean water and sanitation.				X
7. Affordable and non-polluting energy.				X
8. Decent work and economic growth.				X
9. Industry, innovation and infrastructures.		X		
10. Reduction of inequalities.				X
11. Sustainable cities and communities.	X			
12. Responsible production and consumption.				X
13. Climate action.		X		
14. Underwater life				X
15. Life in terrestrial ecosystems.				X
16. Peace, justice and solid institutions.				X
17. Partnerships for achieving goals.				X

Table A.1: Project relationship with the SDGs

demonstrates a strong alignment with Goal 11 (Sustainable Cities and Communities), its direct impact on Goals 9 (Industry, Innovation, and Infrastructure) and 13 (Climate Action) may depend on the extent to which the project's outcomes are implemented and scaled up in real-world transportation systems.