Escuela Técnica Superior de Ingenieros Informáticos

Universidad Politécnica de Madrid

Estimation of Origin-Destination Matrix from Smart Card Data of Public Transportation of the Community of Madrid

Master Thesis

Master in Data Science

AUTHOR: Doga Cengiz

TUTOR: Óscar Corcho

Madrid, Julio 2022

Acknowledgements

I would like to thank my family, especially my sisters, for their support and encouragement through my studies. I thank my partner for being always by my side and motivating me to achieve further.

I would like to express my sincere gratitude to my tutor Óscar Corcho for giving me the opportunity to carry out this thesis and for his guidance throughout the process.

I am grateful to have friends accompanying me all these years and facilitating the journey in the worst moments.

Lastly, special thanks to all the instructors of this degree, for sharing their knowledge, which has allowed me to start my professional career fully prepared and made me appreciate the value of the knowledge.

Abstract

The increase in the population of the cities brings the growth in the public transportation demand with itself which requires achieving the most efficient and faultless public transportation management. One way to accomplish this is to forecast the citizen mobility flows and understand the traveling behaviors of the passengers. In this thesis, origin-destination (OD) matrices are introduced to fulfill these objectives for the Madrid metropolitan area.

Origin-destination matrices are constructed by aggregating the origin and the destination of the trips. To generate the OD matrix for the public transportation networks with entry-only automated fare collection systems such as the the one in Madrid metropolitan area implies the need for destination estimation of each trip.

In this thesis, an algorithm is developed by utilizing the trip chaining method to detect the transfers and predict the destination location of the trips performed by the senior citizens of the Community of Madrid. A sample of data belonging to 3 different days is used to prove the algorithm and destinations of around 29% of the trip are estimated. Moreover, the analysis extracted from the obtained OD matrix is demonstrated.

From the findings of the thesis and the trip chaining algorithm created dedicatedly for the Madrid metropolitan area public transportation network, insights for the future development of the public transportation network and additional knowledge such as train loads, crowd analysis, and travel forecast can be acquired.

Contents

List of Figures

List of Tables

1 Introduction

This study aims to generate an origin-destination matrix based on each trip performed by the senior citizens of the Madrid metropolitan area by using the public transportation system. In order to achieve this goal, it is needed to estimate the destination locations of each trip. For this reason, analyses were performed to understand the public transportation mobility behaviors of the senior citizens and the trip chaining method is applied to estimate the destination locations.

1.1 Motivation

Public transportation is the commonly used mobility mode for the big and developed cities all over the world [1]. Reports that were published before COVID-19 by Consorcio Regional de Transportes de Madrid (CRTM) [2] show that the demand for public transportation has predominantly been uptrend, especially in the previous five years as seen in figure 1.1. Madrid is the second-largest city in European Union and the largest one in Spain with a 3.305.408 population, and the Community of Madrid has 6.751.251 inhabitants [1]. Moreover, the population of the Community of Madrid has been increasing over the years (figure 1.2).

The metrics used to evaluate the transportation modes show numerous benefits of using public transits. According to the urban transportation system report published by McKinsey [1], affordability, availability, efficiency, and sustainability metrics are some of the reasons that make residents choose to use public transportation over private mobility options. An increase in the use of public transportation requires optimization of the system management in order to operate with minimum failure and maximum efficiency.

[1] https://www.ine.es/jaxiT3/Datos.htm?t=2881

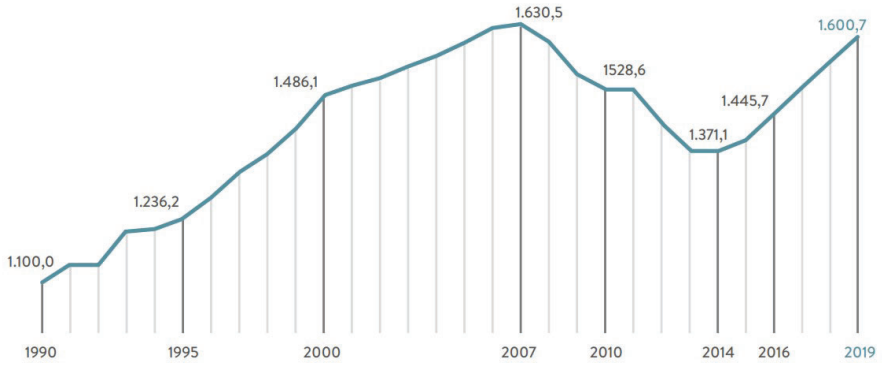

Figure 1.1: Number of public transportation passengers in the Community of Madrid (in millions) by years[2].

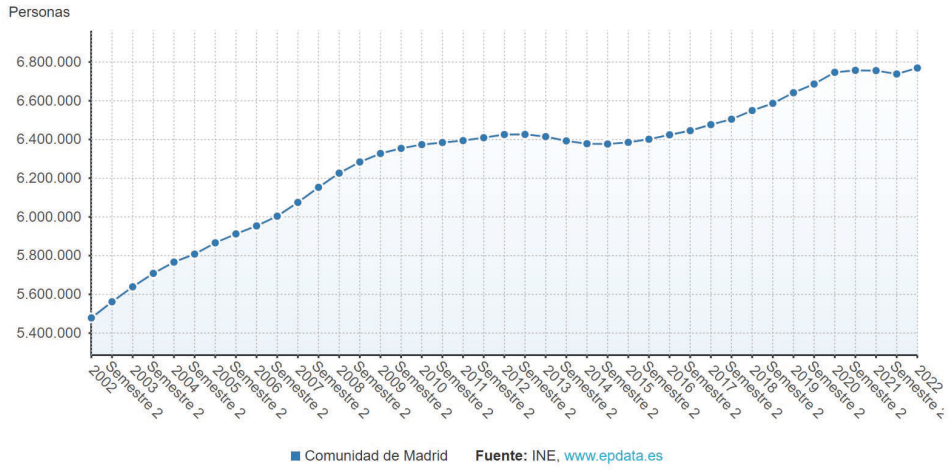

Figure 1.2: Population of the Community of Madrid by years. [2]

Origin-destination (OD) matrices consist of the aggregation of the starting location and the target arrival location information of each trip. Traditionally, for the public transportation domain, these matrices have been generated by using smart card transaction data. Significant inferences such as the comprehending behavior of cardholders, observing the patterns of their trips, estimation of the passenger loads, predicting the passenger traffic flow, and much more can be mined from these matrices. The demand between the locations can be revealed and used for planning the future development and managing the current transportation network.

Automated fare collection (AFC) systems collect valuable data from the smart cards that can be used to estimate the OD matrix and achieve the goals mentioned in the

[2] https://www.epdata.es/evolucion-poblacion/4d9f26fa-ff32-4aad-bf83-7d03c79062ee/madrid/304?accion=1

previous paragraph. A very prevalent AFC system type usually records the data which is created when passengers use their smart card on boarding to start their trips, but not when they are finishing their trips. Therefore, the alighting location has to be estimated in such a system. In this thesis, the trip chaining method is studied to estimate the OD matrices. This method is modified and improved to comply with the characteristic requirements of the Madrid metropolitan area public transportation network.

1.2 Objectives

The main objective of this study is to estimate the origin-destination matrix for Madrid metropolitan area public transportation system which is formed mainly by a large train and bus network. To achieve this objective, the trip chaining algorithm will be deployed. The trip chaining method has many variables that need to be modified depending on the structure of the area. Therefore, the Madrid transportation system will be analyzed to find the most suitable parameters to use in the trip chaining method. The algorithm developed for the trip chaining method will be implemented using Python programming language with Notebooks. The OD matrix obtained from the script can be used for further studies such as predicting the train loads, analyzing the passenger flow, and estimating the demand for each public transportation type.

1.3 Madrid Metropolitan Area Public Transportation System Outline

Madrid metropolitan area public transportation system operates on an 11.000 km network with more than 4 million trips per day[3] [2]. Even though the density of the metro stations and bus stops is much higher in the city center, this transportation network makes it possible to reach any zone within the Community of Madrid which is composed of 30 municipalities (shown in figure 1.3).

[3] https://www.comunidad.madrid/en/inversion/madrid/transporte-publico-excelencia

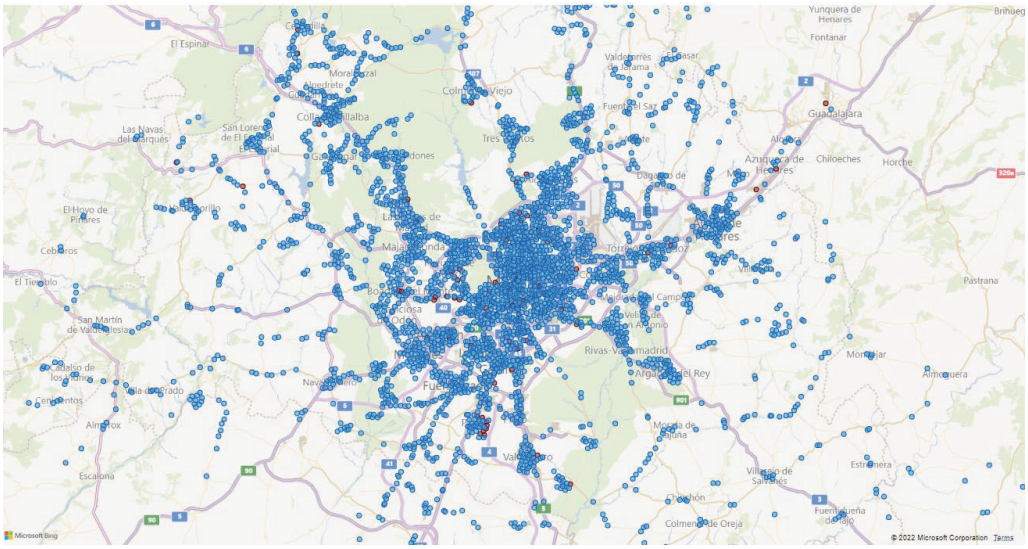

Figure 1.3: Train stations (red) and bus stops (blue) in the Madrid metropolitan area.

1.3.1 Transportation Operators

Madrid's public transportation network consists of 6 different operators. While three of them are operating trains, the rest operates buses for urban and suburban areas;

- **Metro de Madrid:** The Metro de Madrid network comprises a total of 12 lines plus the Branch between the Ópera and Príncipe Pío stations, covering a length of 269.5 km and a total of 236 network stations, of which 39 are multiple-line (with connections between 2 or more lines), and 197 are single-line stations. If counted in terms of lines, the network has a total of 286 stations-line [2].

- **Urban buses of Madrid (EMT):** Empresa Municipal de Transportes de Madrid (EMT) operates the Madrid urban bus network which has 212 lines, 185 of these lines operate during the day and 27 of them during the night lines. Moreover, there is a 24-hour Airport Express line [2].

- **Urban buses of other municipalities:** This bus network of the road transport concessions of the Autonomous Region of Madrid serves in the other municipalities than Madrid municipality [2].

- **Suburban buses:** Suburban buses travels between the different municipalities [2].

- **Light rail (Tramway)**: There are 4 lines that serve on the light rail. They are ML1, ML2, ML3, and ML4 [2].

- **Renfe Cercanías (Suburban train)**: Renfe-Cercanías Madrid is a rail service that connects the city of Madrid and its metropolitan areas, as well as the region's major densely populated areas with the city of Guadalahara. Madrid's Renfe Cercanías Network operates a total of nine routes over 391 km with two branch offices and 95 network stations [2].

In table 1.1 the number of lines and the number of the stations/stops per the transportation type are shown.

	Metro	EMT Buses (Urban buses)	Suburban and Interurban buses	Tramway	Cercanías
Nº of lines	12+Ramal	211	459	5	9
Length-network (km)	269,5	1.598	8.614	54,8	391
Length-lines (km)	269,5	3.856,8	21.271	54,8	713,8
Nº of stations /network stops	237	4726	8349	62	95
Nº of stations /line stops	287	11074	21797	63	178

Table 1.1: Line and station numbers per transportation type [2].

According to the annual report of 2019 published by CRTM, 42,3% of the total trips are made via metro which is followed by EMT urban buses with 27,5% share [2]. Moreover, while demand for rail-based transportation, metro, Renfe Cercanías, and tramway have been increasing over the years, the demand for buses is decreasing (figure 1.4).

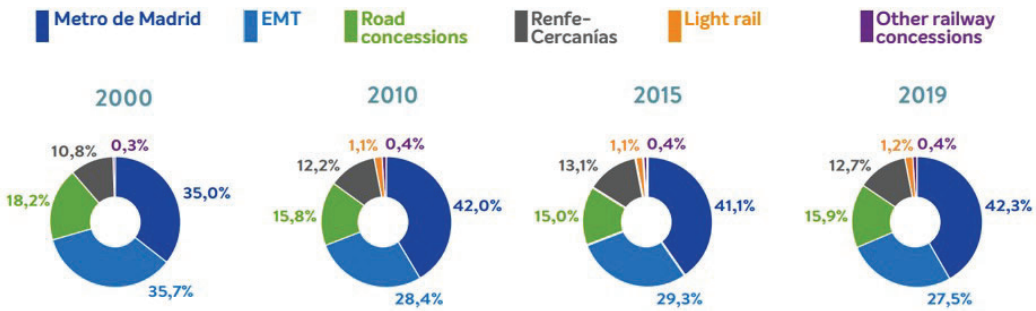

Figure 1.4: Demands for different transportation operators by years [2].

When we check over the metro of Madrid section of the CRTM report, we can see that line 6 is the most demanded line with almost 112 million rides in 2019, and line 11 is the least used line with around 5.5 million passengers in the same year. On the other

	Metro and tramway	Cercanías	Interurban buses	Urban buses
Atocha	1	8	3	26
Nuevos Ministerios	3	7	-	11
Sol	3	2	-	2
Avenida de America	4	-	12	11
Moncloa	2	-	48	19

Table 1.2: Number of transfer options by stations [2].

hand, the most visited five stations are respectively Sol, Moncloa, Nuevos Ministerios, Principe Pio, and Plaza de Castilla.

Other statistics show that Atocha, Nuevos Ministerios, Sol, Avenida de America, and Moncloa are the intermodal points where most transfer are made by passengers [2]. Renfe Cercanías involves the majority of the transfers in Atocha, while for the rest of the connection points the main operator is the metro. Atocha is the most important station for Renfe Cercanías, 8 out of 9 Renfe Cercanías lines start or pass through this station, additionally, 26 urban buses have stops in Atocha. Nuevos Ministerios and Chamartin stations follows Atocha with 7 lines of Renfe Cercanías. On the other hand, in Moncloa, 48 interurban and 19 urban buses operate. For more information on the operator type and line numbers for the most preferred connection points check the table 1.2. We can say that a big percentage of the passenger who travels to these stations intend to have a transfer instead of aiming to arrive at their destinations. Also, it can be seen that station Sol have a relatively lower number of transfer options, however, it is the 3rd most used station for transfer.

1.3.2 Madrid Transportation Zones

The transportation system area is divided into 8 zones of tariffs. 6 of these zones are in the Community of Madrid and 2 of them (E1, E2) are in Castilla-La Mancha [4]. Depending on the passenger profile and ticket type, the price to travel to a zone changes. The zone map can be seen in figure 1.5, zone A covers the center of Madrid while getting further from the city center the zone codes go forwards in alphabetical order. The zones from A to C2 are nested which means they enclose the surroundings of the previous zone. Therefore, a ticket paid for any zones between A to C2 is valid for their inner zones. For example, a ticket for B2 is valid for zones A, B1, and B2. However, the rule for zones E1 and E2 is different. While the zone E1 ticket allows the passenger to travel in zone E1

[4] urlhttps://www.crtm.es/billetes-y-tarifas/zonas-tarifarias.aspx?lang=en

and the Community of Madrid, the zone E2 ticket allows to travel in zone E2 and the Community of Madrid.

Figure 1.5: Madrid public transportation zones [2].

1.3.3 Automated Fare Collection System

At every metro station, and on the buses, there are fare collection machines where users can tap their transportation cards and pay the travel fee to have an access to the vehicle. For each transaction, these automated fare collection (AFC) machines read data from the smart card and record it to the CRTM database. Examples of these collected data are the card serial number, transaction time, user profile code, etc.

The AFC system in Madrid is entry-only, which means the passenger only needs to tap their cards in order to start their trip. This system is also called an open system. There are many cities such as New York City, Istanbul and Santiago de Chile that use entry-only kind of AFC. Another system is called entry-exit, also known as a closed system, which requires users to tap or swap their cards to leave the vehicles or stations. Seoul in Korea and SEQ in Australia are examples of the cities that implemented this system in their public transportation. London transportation system has a bus network with open and underground network with a closed system.

In Madrid, except for a few stations, such as the Cercanías station, there is no AFC machine on exits because the AFC system was designed as entry-only. The dataset available for this research contains information about the starting point of a trip, however, no information is available on the destination. Moreover, passengers do not need to pass through the fare collection gates in the case of metro transfers in the same station and because of this, metro transfers are not recorded hence should be inferred.

1.3.4 Smart Card Types, Profiles, and Tickets

Madrid public transportation card is used for all public transportation types: metro, urban and suburban buses, light rail, and suburban railway (Cercanías Renfe). Users can purchase a different kind of tickets based on the preferred zones and depending on the purchased zone range price of the tickets change.

Transportation cards can be personalized to be eligible for discounts. For example, users younger than 26 can have a young card monthly or annually subscription that has a fixed price regardless of the zones, and it can be used on any trip between zone A-E2. Similarly, an elderly card monthly subscription that is valid from zone A to C2 can be obtained by users older than 65 years old.

According to the CRTM report 2019, there are 4.374.771 active personal smart card users which are 10,4% higher than the previous year. From the most common profile to the least, there are respectively, around 2 million regular, 1,224 million young, almost 900 thousand elderly, 87.8 thousand kids, and 83.5 thousand blue card profiles. Additional to personal cards, there are around 12 million non-personal MULTI cards were in circulation during 2019. Moreover, the report shows that the majority of the trip payments, 78,1% of total trips were done by smart card with a transportation pass. The 10-trip tickets and one-way tickets follow the transportation pass respectively [2].

Smart card types

There are four different types of public transportation cards, two of them can be obtained by everybody, yet the other two require certain criteria to own.

- **Multi Card:** A multi-card is a 10-year multi-personal, rechargeable, contactless public transport card that can be used in combination with various types of tickets.
- **Public Transport Personal Card:** This is a personal, non-transferable card that expires 10 years from the date of issue and can be used to replenish both personal and non-personal tickets. The card contains the cardholder's name and photo, as well as the card identification number.
- **Children's Card:** Children's public transport cards are free and are intended for children ages 4-6. With this card, children get free access to all public transportation in the Madrid Autonomous Region for the life of the card.
- **Blue Card:** This transportation card is similar in function to a personal public transportation card that can only be used with the blue card ticket type of the same name. This card is intended for citizens registered in Madrid who meet certain age or disability requirements and have limited funding. Can only be used for travel on Metro de Madrid (Zone A), EMT, light metro route 1 (ML1).

Smart card profiles

The personalized smart cards hold one of the following profiles based on the age of the users. The transportation fares vary depending on the profile of the card.

- **Young (Abono Joven):** 26 years old and younger passengers.
- **Normal (Abono Normal):** 26-65 years old passengers.
- **Senior (Abono Tercera Edad):** 65 years old and older passengers.

Apart from the type of the profiles, every user can have the following discounts on the fees if they meet the conditions.

- **Big family - general:** 20% discount
- **Big family - special:** 50% discount
- **People with disabilities bigger than 65%:** 20% discount

Ticket types

There are different types of tickets that were designed for the different traveling expectations of the passenger who can purchase in accordance with their needs.

- **Single ticket:** This type of ticket is valid for all operators. This ticket allows the passenger to have one trip and it was designed mainly for occasional users who buy the ticket in the moment of travel.

- **10-trip ticket:** Passengers who purchase this ticket can travel 10 times with it. This type of ticket is valid for all EMT lines, in zone A of the Madrid Metro network and on Light Rail route 1.

  - Metrobús: Valid for EMT, Metro Zone A, and (Madrid light rail tram route 1 (ML1).
  - Bus+Bus: Valid for EMT with change.

- **Travel pass:** Only can be utilized with personal transportation cards. There are two different passes, monthly 30-day passes and annual passes which have unlimited use during these periods for all operators. The price of the travel passes changes based on the user profile.

1.4 Methodology and Reproducibility

1.4.1 Methodology

The data used in this study is provided by the CRTM, and it consist of two different kinds of dataset.

- **Smart Card Transaction Records:** Dataset that stores all the transactions performed by the Madrid metropolitan area public transportation smart card users with the senior profile. The records from 01-01-2016 to 31-02-2020 are available, however, for this study, only the data from the first week of February 2020 is used.

- **Public Transportation Topology:** Dataset that contains all the stops and stations with their location and service information.

1.4.2 Reproducibility

- The code used in this study is publicly available in the GitHub repository [5].
- The data used in this project is not publicly available due to the regulations to protect the data and the privacy of the smart card holders.

1.5 Thesis Organization

This thesis is organized into 6 chapters. In chapter 2, the previous studies on public transportation mobility and especially on origin-destination matrices are presented. The assumptions of the OD matrices were explained in detail as well as the validation strategies. Chapter 3 describes the datasets used for the thesis and the preprocess applied to these datasets to prepare them for the further steps. In chapter 4, the approach to developing an algorithm to generate the OD matrices is explained. Chapter 5 presents the results obtained by applying the algorithms described in the previous chapter. The OD matrix generated and the other analysis obtained from the OD matrix are shown for different time periods. The first section of the final chapter summarizes the performance of the algorithm and the findings of the thesis. The last section suggests the potential future works that can improve the present state of this thesis.

[5] https://github.com/dogacengiz/TFM

2 Literature Review

The literature review chapter has been divided into two sections. In the first section, various studies conducted on public transportation forecasting and analysis are mentioned briefly, and in the second part, the state of the art of origin destination matrices is explained in detail.

2.1 Public transportation mobility studies

When we look at the research papers on public transportation, the goal of some of these studies generally emphasizes the importance of optimizing the service, obtaining better efficiency, exploring the behaviors of the passengers, and increasing the service quality. To achieve these goals, predicting the demand for a specific transportation mode at a certain location and time is crucial. If a method can estimate the passenger load for a certain situation with high accuracy, many other important inferences can be fetched out of the user travel dataset.

Although the methods change subject to the available dataset, we can see many machine learning methods, especially deep learning-based ones, applied in this domain. There are examples of using support vector machines, random forest and gradient boosting decision trees for short-term demand prediction, and usage of classifiers to identify the passenger load category [3][4][5]. Also, the combination of recurrent neural networks and convolutional neural networks is utilized by many researchers for capturing spatial and temporal inferences. Additionally, the application of the statistical approaches that are associated with time series exists.

Recurrent neural network (RNN) is one of the most preferred artificial neural network algorithms. It is known for its capability to display the temporal dynamic behavior of the data by using the internal state memory of the units. This feature makes RNNs suitable for their applications on a sequence of data. However, RNNs suffer from the vanishing gradient problem which means the update of the gradient gets exponentially

smaller, and the learning stops. Therefore, RNNs are not good enough for inferring the long-term dependencies in the time series.

Long-short time memory units solve the vanishing gradient problem by adding three gates to the structure of the units. These gates are input, output, and forget gates that control and update the cell states. LSTMs encode the long-term dependencies into the cell state vectors thereby the vanishing gradient problem is prevented. Passini [6] used long-short time memory (LSTM) units in a neural network to forecast the univariate train loads for each station in the northern area of suburban Paris by using the schedule of the train, sensors in the train that count the number of the passenger, and calendar data. He modified the LSTM units by adding an encoder and decoder to capture the influence of the dynamic context on the series. Toque et al.[7] also deploy different versions of gated recurrent units (GRU) for forecasting the short-term demand of railways during special events.

Ding et al.[8] used a statistical method for modeling the short-term metro ridership forecast by focusing on the dynamic volatility of the data. In their paper, the autoregressive integrated moving average (ARIMA) model was used for the main part of the forecast, while the generalized autoregressive conditional heteroskedasticity (GARCH) model was used for the volatile part which is the forecasting of the residuals.

Two master's final thesis studied the analysis of senior citizens' travel behaviors by using the CRTM data of senior citizens. Lacki [9] performed an explanatory analysis of the mobility characteristics in the frequency, temporal, and spatial manners as well as clustering the users to determine the prevailing groups. The other thesis studied by Puerta [10] aims to discover main the characterization and prediction of the individual's use of public transport in the Community of Madrid by people over 65 years of age.

2.2 Origin-destination Matrix

Origin destination matrices represent the individuals' mobility in a certain geographic space from the origin (O) to destination (D). OD matrices demonstrate the flow between two geographic points, zones, or transportation stations. Once OD matrices are generated further analysis of public transportation and human flow can be done. However, due to the type of the public transportation structure, and the station and vehicle hardware, obtaining the OD matrix may be challenging.

2.2.1 OD Matrix Estimation Methods

There are different kinds of approaches to estimating the OD matrices depending on the available data due to the type of the automated fare collection system of the public transportation network.

Trip Chaining Method

The trip chaining method was introduced by Barry et al. [11] in 2002 to estimate the OD matrix for New York City public transportation trips by using the data from the entry-only automated fare collection system. MetroCard is the smart card used in New York City for metro and bus trips, and it records the card ID number, code of the boarding station, and date for each time the card is swiped on an AFC device. The biggest challenge of generating the OD matrices for an entry-only system is to infer the alighting station. Barry et al. [11] proposed two assumptions based on public transportation travel behaviors that were tested by using travel survey data collected by the transportation council, and after validation, results showed that these assumptions are correct for 90% of the trips according to the same survey data.
The assumptions of the work of Barry et al.[11] are:

- A high percentage of riders return to the destination station of their previous trip to begin their next trip.
- A high percentage of riders end their last trip of the day at the station where they began their first trip of the day (known as the day's symmetry trip assumption).

The idea behind the second assumption is that the users start their first trip somewhere close to their home, and eventually their last trip aims to arrive at home.

Figure 2.1 demonstrates the trip chaining method. B1 represents the first boarding location where in the buffer zone around the B1 the resident of the passenger is expected to be located. A1 is the alighting stop of the first trip. If there is no transfer from the first trip leg to the second trip leg, during the weekday, A1 is expected to be the working location for adults and educational institutions for young cardholders. In other cases, A1 is a location that passenger goes for an activity. The alighting point of the last trip leg, A3, is supposed to be somewhere close to the first boarding location, B1. If the passenger performs a transfer the successive alighting and boarding locations should not be further than the maximum transfer distance (MTD).

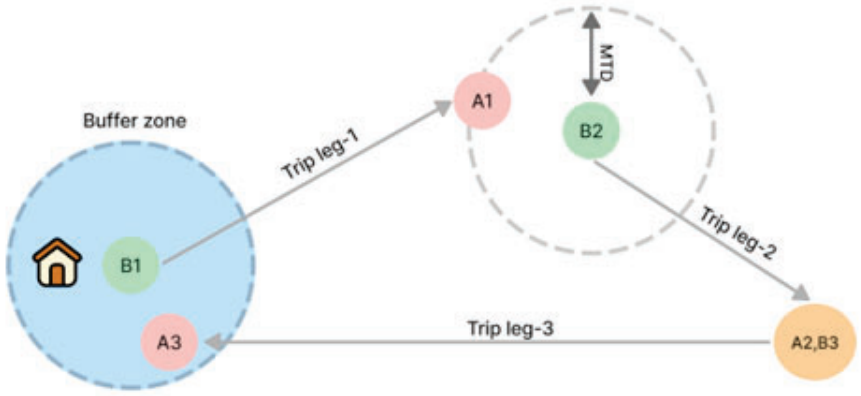

Figure 2.1: Trip chaining method illustration.

Two years later, J. Zhao [12] enhanced the assumptions proposed by Barry et al. [11] by suggesting the following three hypotheses:

- Traveler does not use other types of private transportation modes such as car, bicycle, or motorcycle between two successive public transportation trips.
- There is a limit on the distance that a traveler walks for transfer. (In Zhao's studies this distance was taken as 1320 feet, equivalent to 401 meters.)
- The last destination of the day should be the same station as the first trip of the day.

The last two assumptions are used widely in origin-destination matrix research.

After three years, in the work of Trépanier et al. [13] the day's symmetry trip assumption (formulated by Barry et al.[11]) was relaxed, as these stations do not have to be

In the last years, many different individual private transportation types came into the metropolitan residents' life such as shared micro-mobility (bicycles, e-scooters), and shared mobility (car-sharing, ride-sharing). These transportation modes make it more possible to fail the first assumption of J. Zhao [12]. According to the 2020 report of Lime, a shared e-bike and e-scooter company, around 9% of the trips were done by the users

for transit access[1]. Moreover, in the case that the user prefers to use any of mentioned private transportation options, the data to track these rides are not available.

The second assumption of J. Zhao [12] requires two parameters, walking distance and walking time. Walking distance is defined as the maximum distance that passengers can consider to walk to have a transportation transfer. In Madrid, the density of the metro and bus stops is very high in the center. All the metro lines are connected which means the passenger can travel in the area around the metro routes without leaving the station, but only changing the platform. CRTM 2019 report [2] shows that there are 4726 EMT urban bus stops in Madrid in addition to the metro stops, therefore in Madrid city center the walking distance to have a transit can be assumed to be very short. According to the statistics published by Moovit, mobility as a service provider and journey planner app, citizens of Madrid walk on average of 650 meters per trip [2]. Different numbers for walking distance chosen by the researchers based on the structure of the cities can be seen in the table 2.1. Also, some researchers set a maximum transfer time to be spent by the passenger to change the transportation vehicle (Table 2.2).

Distance (m)	Study	City
402	Zhao et al., 2004 [12]	Chicago, US
530	Alsger et al., 2016 [14]	SEQ, Australia
640	Nunes et al., 2016 [15]	Porto, Portugal
750	Gordon et al., 2013 [16]	London, UK
800	Alsger et al., 2015 [17], Nassir et al., 2011 [18]	SEQ, Australia London, UK
1000	Yan et al., 2019 [19], Munizaga et al., 2012 [20], Wang et al., 2011 [21], Munizaga et al., 2014 [22]	London, UK Santiago, Chile Shenzhen, China
1100	Cui., 2006 [23]	Chicago, US
2000	Trépanier et al., 2006 [13]	

Table 2.1: Maximum transfer distances used in the specialized literature [24][25].

[1]https://www.li.me/blog/limes-2020-wrap-up-report-highlights-turning-point-for-micromobility
[2]https://moovitapp.com/insights/en/Moovit_Insights_Public_Transit_Index_Spain_Madrid21

Time (min)	Study
18	Barry et al., 2009[26], Ali et al., 2016[27], Bagchi et al., 2005[28]
30	Liu et al., 2019 [29], Munizaga et al., 2014 [22], Munizaga et al., 2012 [20]
60	Alsger et al., 2016[14], Alsger et al., 2015[17], Mosallanejad et al., 2019 [30], Nassir et al., 2015[31], Yan et al., 2019 [19]
90	Hofmann and Mahony, 2005 [32], Kumar et al., 2018 [33], Nassir et al., 2011 [18]
Variable	Chu and Chapleau, 2008 [34], Gordon et al., 2013 [16], Seaborn et al., 2009 [35], Yap et al., 2017 [36]
<transit frequency	Huang et al., 2020 [37]

Table 2.2: Maximum transfer time used in the specialized literature [24][25].

The trip chaining algorithm estimates the destination of the first trip based on the boarding station of the next trip. Therefore the algorithm has a deficiency in estimating the destination of the trips in case there is only one trip made by the user during that day. When there are no consecutive trips, it is not possible to create chains and predict the alighting of the trip. When some researchers decided to remove these single transactions [14], other authors tried to overcome this problem in different ways. Trépanier et al. [13] analyzed the similar activities of the same user on the other days to estimate the single transaction destinations. In most of the studies, [15][20][21], the single transactions were mentioned as one of the reasons of higher error. Besides the single transaction issue, the other common problems quoted are listed below:

- Passengers walk more than the maximum distance.
- Mistakes in data such as missing values and duplicated transactions. The standard solution for these errors is to eliminate the row.
- Passengers whose last trip's destination is not the same as the initial point of their travels.
- Usage of non-pubic transportation modes between two transactions of smart card.
- The distance between the successive alighting point and boarding point may be longer than assumed.

To apply he day's symmetry trip assumption and to remove the single transactions in a day, the beginning and the end of the day should be defined. In general, the actual

midnight 0:00 AM is not the time that activities on public transportation reach the least. According to the OD matrix in Santiago, Chile described in the studies of M. Munizaga et al. [22], after 0:00 AM, there are activities belonging to the end of previous day trip chains. For this reason, the hour that has the least amount of transactions, which is 4:00 AM for Santiago, was detected and used as the virtual midnight. This change reduces the error and the number of ignored trips.

Other Approaches to Estimate the OD Matrix

There are other methods less widely used than the trip chaining method for estimating the OD matrices such as probabilistic and deep learning-based models.

The first probabilistic approach was proposed by Dou et al. [38] that calculates the probability of alighting station for the remaining stations by taking into account the travel distance and number of passengers. Later, this model was enhanced by other researchers by considering the capacity of the stations and land use grades nearby the stations [39][40][41]. This approach focuses on the total in-out passenger number per station, however, it does not allow to infer the individual alighting station.

Another model created to estimate the OD matrices is a deep learning-based model. Artificial Intelligence models are widely used for traffic predictions, metro train load forecasting, and many other tasks. In the public transportation domain, datasets are usually a time series, and deep learning algorithms such as recurrent networks, LSTMs (long-short time memory), and CNNs (convolutional neural networks) are used to achieve the goals mentioned.

Jie [42] used a back propagation neural network algorithm to estimate a OD matrix for travels via buses. Another study was conducted by Jung et al. [43] by using the smart card data and the land use information to estimate the alighting point.

Deep learning models have one significant shortcome that the dataset needs to contain the alighting location information. However, most of the transportation AFC systems, including the ones in Madrid metropolitan area, are built as entry-only. Deep learning is supervised learning which means the model learns from the dataset. The model modifies the parameters to improve the output estimations to converge to the real output. With an entry-only system data, the model cannot be trained, therefore this approach is not appropriate to predict the alighting location. Another handicap of deep learning models are closed box system and the reason for the output values cannot be observed.

2.2.2 Validation Methods

Validation of the proposed methods for alighting stop estimation is one of the biggest difficulties. Due to the lack of alighting station information for the only-entry systems, the estimation accuracy cannot be tested and validated directly.

There are two different types of validation techniques; endogenous and exogenous validation. Endogenous validation is made by using the same dataset or a part of the same dataset that is used for developing the algorithm. Contrary to this, exogenous validation uses other external datasets to validate the proposed algorithm. The algorithm created for entry-only systems can be validated by using any of these options. However, due to the existence of alighting station information in entry-exit systems, using only the endogenous method is adequate. In our case, there is no other available source than the dataset used for development, therefore endogenous validation method is appropriate.

Until 2013, researchers preferred either to use exogenous validation methods or not validate the algorithms. Barry et al. [11], used the travel diary survey records supplied from NYMTC. 590 trips were chosen by omitting the cards that have only one transaction in a day. This small set validation showed that 90% of the estimations match the survey data. In the same paper, it is also mentioned that the turnstile located on the exit gates of each station counts the number of the leaving passengers, however, it does not record any other information because smart cards are not required to leave by the turnstiles. This count number can be used to control the OD matrix by not confirming the individual alighting points but by checking the total number of the leaving passengers. However, the count number is not absolutely accurate because passengers can use the emergency exit gates.

A similar validation technique was used by Zhang [44]. A customer survey record of the Chicago Transit Authority CTA was used for partial exogenous validation. Another study conducted by Farzin [45], used survey data from 1997. Due to the time gap of over 10 years, the validation of the study is not considered convincing.

A list of the validation methods used in literature is given in table 2.3. This summary of the literature was obtained by reviewing two research papers which can be read for further detailed information [24][25].

Authors	Validated propose	Validation method	Validation dataset
Barry et al. (2002) [11]	Trip chaining assumptions	Exogenous	Travel diary survey
Cui (2006)[23]		No validation	
J Yu et al. (2006) [42]		Exogenous	Bus survey data
Dou et al. (2007) [38]		Exogenous	Bus survey data
Zhang Lianfu (2007) [44]		No validation	
Zhao et al. (2007) [46]	tOD matrix	Exogenous / Partial validation	CTA Customer OD survey
Barry et al. (2009) [26]	Multi-modal OD inference model	Exogenous	Entrance and exit counts in the subway, ride check data for bus
Farzin (2008) [45]	Zone-to-zone bus OD matrix	Exogenous	OD Household Survey
Seaborn et al. (2009) [35]	Multi-modal OD inference model	Exogenous	LTDS
Nassir et al. (2011) [18]		No validation	
Wang et al. (2011) [21]	Inference of bus OD matrix	Exogenous	BODS[a]
D Li et al. (2011) [47]		No validation	
Munizaga et al. (2012) [20]		No validation	
Gordon et al. (2013) [16]	Inference of OD matrix for intermodal	Exogenous	LTDS[b]
Zhang et al. (2014) [41]		Exogenous	Bus survey data
Munizaga et al. (2014) [22]	OD estimation and transfer algorithm	Exogenous	OD metro surveys
Alsger et al. (2015) [17]	Transfer walking and time, and days symmetry assumption	Endogenous	SCD[c] (entry-exit system)
He et al. (2015)	Alighting location estimation	Endogenous	SCD (entry-exit system)
A Nunes et al. (2016) [15]		No validation	
Nassir et al. (2015) [31]	Transfer detection	Exogenous	HTS
Alsger et al. (2016) [14]	Transfer walking and time	Endogenous	SCD (entry-exit system)
Jung et al. (2017)	Alighting location estimation	Endogenous	SCD (entry-exit system)
Kumar et al. (2018) [33]	OD estimation	Exogenous	On-board surveys
Liu et al. (2019) [29]	Transfer detection	Endogenous	SCD (trip-chaning)
Yan et al. (2019) [19]	Alighting location estimation	Endogenous	SCD (entry-exit system)
Huang et al. (2020) [37]	Inference of bus OD matrix	Exogenous	Ticket recycling survey
Egu and Bonnel (2020) [48]	OD estimation, transfer algorithm	Exogenous	HTS and onboard OD surveys
Assemi et al. (2020) [49]	Alighting location estimation	Endogenous	SCD (entry exit system)

[a] Bus Origin Destination Survey
[b] London Travel Demand Survey
[c] Smart Card Data

Table 2.3: Validation methods used by literature [24][25].

3 Data Description and Preprocessing

In this chapter, the datasets used to verify the proposed trip chaining method are described with example data as well as the 24-hour sample transaction data selected. In the last section, the data preprocessing methods applied to datasets are explained briefly.

3.1 Data Description

To perform this thesis, two different types of datasets were provided by the CRTM, transaction records of the Madrid metropolitan area public transportation smart card, and the network topology of the public transportation system.

Another additional dataset was created manually to link together the train lines with their stations.

The provided datasets consist of many columns which are not necessary for the implementation of the trip chaining method. Below, the utilized columns per dataset are given with their brief definitions.

3.1.1 Smart Card Transaction Data

As mentioned in subsection 1.3.3, in the Madrid metropolitan area public transportation the entry-only system is in use. Therefore, the dataset is formed by the smart card information of the passenger as well as the data of the boarding transportation mode. The transaction records dataset is limited to the records of the users with senior smart card profiles. An example data can be seen in table 3.1.

- **CARDID**: This attribute is the hashed version of the smart card serial number. This identification number is unique for each card. The CRTM provided the dataset

with the encoded card IDs and the hash function is not given in order to avoid any privacy and confidentiality of users.

- **DATE**: The date of the transaction with the format of year-month-day hour-min-sec.
- **DPAYPOINT**: The code of the payment point device. (Same as the DPAYPOINT column of the train stations, and bus stop tables.)

CARDID	DATE	DPAYPOINT
F606DFF24C329924DFFEF79920D5933C	2020-02-01 14:11:42	02_L11_P2

Table 3.1: Transaction dataset example record.

3.1.2 Public Transportation Network Topology Data

Two datasets provided for the network topology are the train station information and the bus stop information datasets. The other dataset created manually demonstrates the relation between the train stations and the train lines.

Train station dataset

Each row in this dataset contains information on each train station that is on tramway, Cercanías, or metro network. Five columns of this table are used (shown in table 3.2). The DENOMINAPARADA column is unique which means there are one row for each station and there are 429 rows in the dataset.

- **CODIGOMUNICIPIO**: The identification code of the districts.
- **DENOMINAPARADA**: The name of the train station.
- **DPAYPOINT**: The code of the payment point device.
- **LATITUD**: Latitude of the train station.
- **LONGITUD**: Longitude of the train station.

CODIGOMUNICIPIO	DENOMINAPARADA	DPAYPOINT	LATITUD	LONGITUD
28079	ABRANTES	02_L11_P2	40.38083	-3.72790

Table 3.2: Train station dataset example record.

Bus stop dataset

This dataset contains information on all the bus stops in Madrid metropolitan area (see table 3.3). Contrary to the station dataset, in this dataset, there are multiple rows containing the information of the same bus stop with different stop-line pairs. The number of rows drops from 36.391 to 34.560, when the records with null IDLINEA values are removed. Moreover, with the elimination of the repeated stops, 13.018 unique bus stop instances are obtained. Five columns of this table are used.

- **CODIGOMUNICIPIO**: The identification code of the districts.
- **DPAYPOINT**: The code of the payment point device.
- **LATITUD**: Latitude of the bus stop.
- **LONGITUD**: Longitude of the bus stop.
- **IDLINEA**: The identification number of the bus (bus number).

CODIGOMUNICIPIO	DPAYPOINT	LATITUD	LONGITUD	IDLINEA
28079	3B_L725_P7227	40.86041	-3.69962	725

Table 3.3: Bus stop dataset example record.

Metro-line dataset

The train station dataset gives details on the location of the station, but not the lines that stop by that station. Because of that, a dataset that contains the station name and a list of line numbers for that station was generated by using the websites of metro de Madrid[1] and Renfe Cercanías[2].

- **stationName**: The name of the train station.

[1]https://www.metromadrid.es/en/linea/linea-1
[2]https://www.renfe.com/es/en/suburban/suburbanmadrid

- **lines**: A list of metro, tramway or Cercanías train line numbers.

stationName	lines
PINAR DE CHAMARTÍN	1;4;ML1

Table 3.4: Train station-line dataset example record.

3.2 Sample data

The dataset used for this thesis consists of the records of senior citizens who are older than 65. Two days, Saturday, February 1, and Tuesday, February 4, were selected to prove the proposed algorithm. Even though the trip patterns of the senior citizens are not explicit as the rest of the population such as students and workers, one weekday and one weekend were chosen to explore the difference in the performance of the algorithm. While February 1 contains 441.086 rows, there are 709.170 transactions recorded on February 4.

3.3 Data Preprocessing

Data processing techniques are applied to the datasets to prepare them for the implementation of the trip chaining algorithm. After the stationInfo and stopInfo files are read into data frames, the rows that have IDPARADA and DPAYPOINT values missing are deleted. As mentioned before stops in the stopInfo dataset are not unique, therefore the stops with the same longitude and latitude are eliminated. These two data frames are concatenated to obtain one data frame called allStopsDf with all the stops and the stations in Madrid metropolitan area. Due to the duplication of the stop ID values of the different public transportation operators, the index of the data frame is used as the IDPARADA (IDSTOP) value.

The smart card transaction records are read to a data frame and a defined 24-hour time slices are filtered. The transaction records that are the single record of the corresponding card ID are removed because they cannot be used for the trip chaining method. Also, the duplicated records are removed.

Similar to the metro-line dataset mentioned at section 3.1.2, a data frame that contains a list of the lines that pass through each bus stop is created from the stopInfo dataset.

4 Trip Chaining Algorithm

The proposed trip chaining algorithm is constructed in three phases. In the first phase, the transfers performed by the passengers are attempted to detect. After, the destination of the trips is estimated. Finally, the OD matrix is generated by using the obtained information from the previous steps. This chapter explains the approaches to implementing these phases.

4.1 Near Stop Finding

A function was created by using the allStops data frame to generate the nearStop table. The function runs a nested loop that iterates through the allStops data frame twice. For each bus stop or train station recorded in the data frame, all the other rows are checked whether they are in a circular area defined with a specified radius distance. As mentioned in section 2.2.1, the average distance that a passenger wishes to walk to have a transfer is 650 meters in the Madrid metropolitan area. Thereby, the distance is taken 650 meters to find the near stops. This table is used later for the transfer detection algorithm.

Numba[1] library is used by vectorizing the inputs as NumPy[2] arrays to accelerate the runtime of this function.

In figure 4.1, the stops in 650 meters radius area of Sol metro station are shown. As it can be seen there are 7 metro stations (red bubbles) and 57 bus stops (blue bubbles) that were found close to Sol metro.

[1] https://numba.pydata.org/
[2] https://numpy.org/

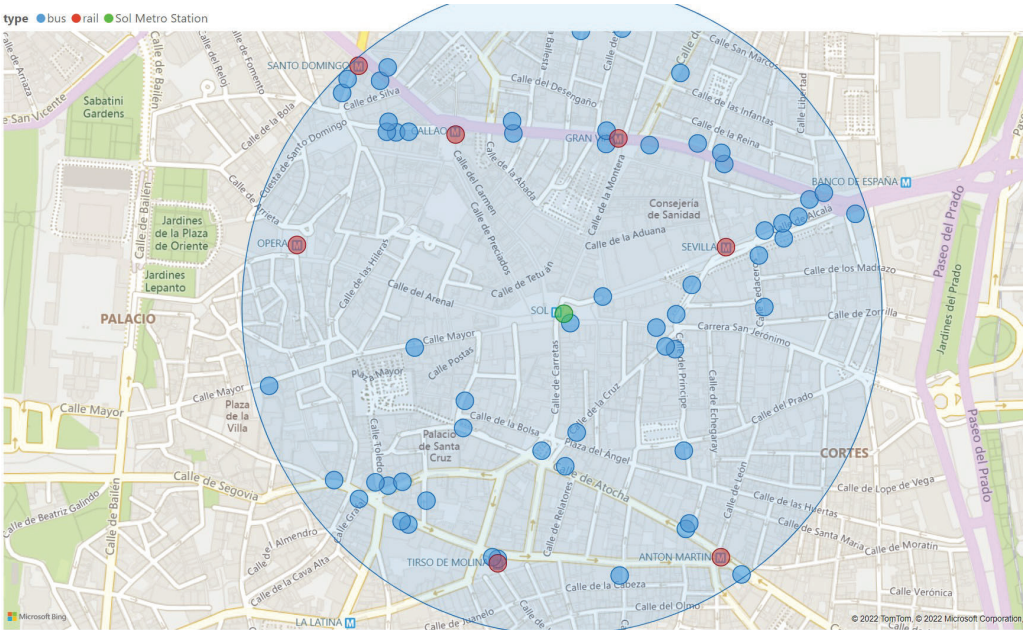

Figure 4.1: Sol metro station near stops.

4.2 Transfer Detection

Each ride made via public transportation has a starting and ending point, however, the end of a particular ride does not mean it is the destination of the passenger. In big cities, it is very typical to have multiple rides with transfers to arrive at the target location. The principal transportation modes in Madrid allow users to have four main types of transfer, these are;

- Train-to-train
- Bus-to-bus
- Train-to-bus
- Bus-to-train

Three types of trains, metro, tramway, and suburban trains (Renfe Cercanías), operate in the Madrid metropolitan area. These different kinds of trains share many joint stations that passengers can transfer from one type of train to another type. The transfers that passenger changes the type of the operator of the train requires them to leave the platform through the turnstiles and use their smart card to enter the other operator's platform. For instance, to transfer from a metro line to a tramway line, passengers need

to tap their smart cards into the AFC machines. Therefore, in this kind of transfer, the change in the train type is recorded in the CRTM databases. However, the real challenge is to observe the train transfers within the same operators because these transfers do not ask the passenger to perform any other transaction while changing the train line. Hence, for this type of transfer, only the first transaction location is found in the dataset, yet no information can be extracted directly about the latter legs of the transfer chain if they exist.

On the other hand, this obstacle does not appear for the detection of the bus-to-bus or bus-to-train transfer types. Each bus trip is written individually to the CRTM databases. In the dataset, the DPAYPOINT attribute varies based on the bus number and the stop ID, thereby, the exact get-on location is known for the bus trips.

For each type of transfer, a different algorithm was developed. At first, the pair of consecutive smart card transactions are assigned to one of these four types of transfers by controlling whether the type of the vehicle is a train or bus for two consecutive transactions. Then, according to the type of the transfer, one of these four algorithms is applied to the transaction pairs.

Each algorithm aims to calculate the maximum time needed to travel from the current boarding location to the following boarding location under certain scenarios summarized in figure 4.2.

The time is calculated in minutes by accepting some assumptions such as using a fixed number for the vehicle waiting time and taking the human walking speed constant for every passenger. Actually, the vehicle waiting time may vary depending on the period of the day due to the public transportation vehicle's timetable. A more precise estimation of maximum time can be done by including the bus and train schedules to derive the exact duration of the trips.

When the maximum time is calculated, it is compared to the actual time difference between two consecutive boarding times. If the actual time is less than the maximum time, it means the passenger has not lost time by doing another activity and is directly transferred to the next vehicle. Thereby, the alighting point of the first trip leg can be estimated. Also, we can say that the target destination of the passenger is the last alighting point of the transfer chain.

In the next sections, each algorithm is explained in detail with the equations used for calculating the maximum time.

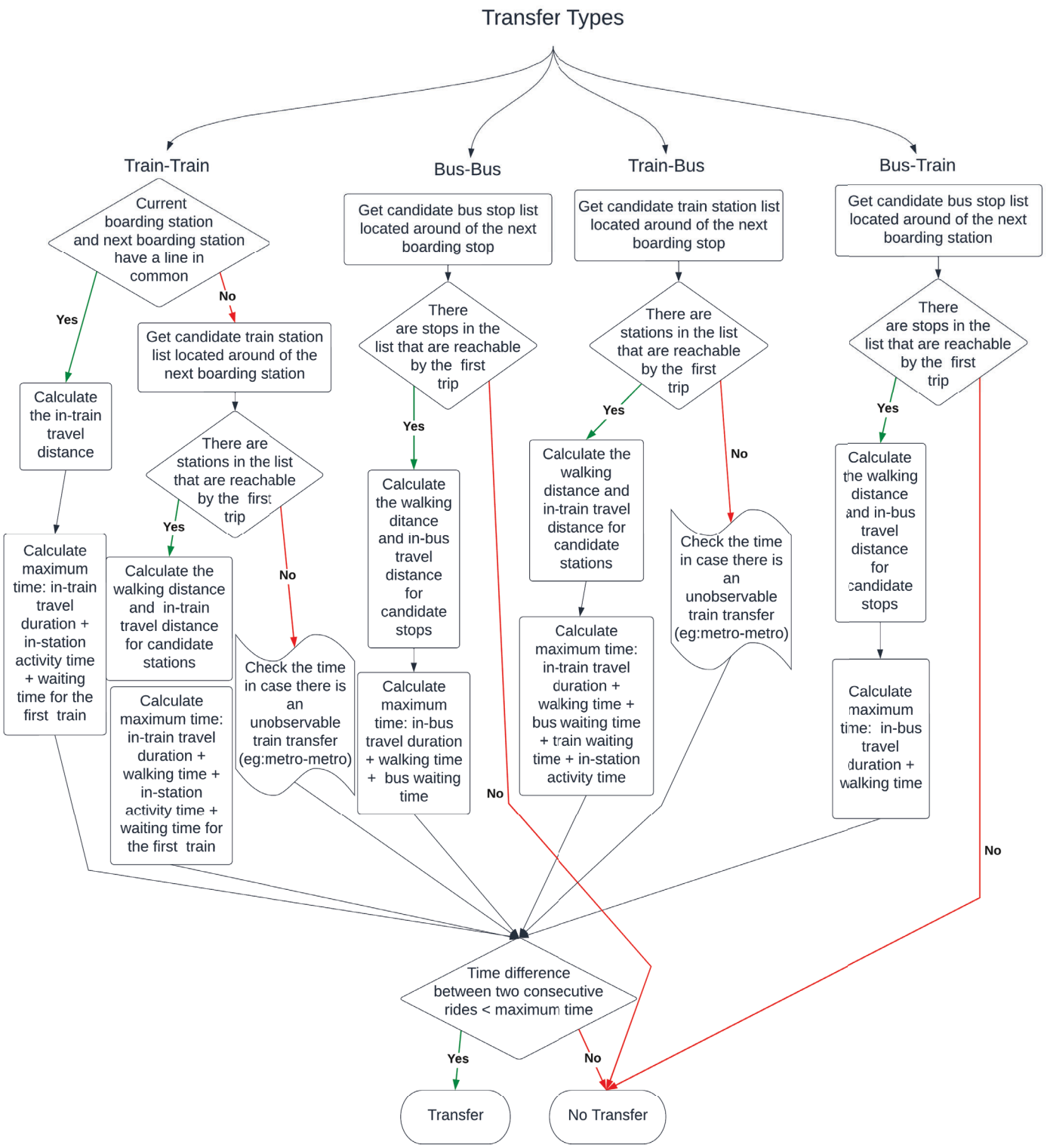

Figure 4.2: Transfer algorithm diagram.

4.2.1 Train-to-train transfer

Firstly, it is checked whether the two consecutive boarding stations have a metro, tramway, or Renfe Cercanías line in common. If they have a mutual line, using another intermediate train to arrive at the second boarding station is not logical. In other words, there is no unobservable in the middle transfer. In the case of the existence of a common line, the ideal traveling time to arrive at the second boarding location is calculated by equation 4.1. The average speed of urban public transportation trains in Madrid (taken as 500 m/min) is used to calculate the travel duration (in minutes) between two boarding locations. Additionally, 5 minutes of in-station activity and Madrid metropolitan area public transportation average vehicle waiting time of 10 minutes[3] are added to the maximum time needed to travel between these two stations.

$$maxTime = \frac{distance(first\_boarding(lat, lon), second\_boarding(lat, lon))}{500} + 5 + 10 \tag{4.1}$$

If these two consecutive boarding station does not have the same train line, the area around the second boarding station is scanned to find the other possible train stations. In this case, we assume that the passenger uses a train first then leaves the station and has a short walk to get access to the next station. Similar to the first case, for each candidate station in the specified area, it is checked if they have a common train line with the first boarding station, if a common line is found, the equation 4.2 is applied to calculate the maximum time. In this formula, there is an additional term to calculate the walking time between the first alighting point and the next boarding station by using the average walking speed of senior people (50 m/min).

$$maxTime = \frac{distance(first\_boarding(lat, lon), candidate\_station\_boarding(lat, lon))}{500}$$
$$+ \frac{distance(candidate\_station\_boarding(lat, lon), second\_boarding(lat, lon))}{50} + 5 + 10 \tag{4.2}$$

The last case is valid for the unobservable same operator transfers during the travel from first boarding to first alighting point. For this case, the maximum time is calculated

[3] https://moovitapp.com/insights/en/Moovit_Insights_Public_Transit_Index_Spain_Madrid-21

by adding extra 10 minutes to the equation (4.1) used for the first case due to the waiting and in-station walk for changing the metro line.

Depending on the case detected, the calculated maxTime is compared with the actual time difference between the two boarding transactions. A transfer is detected when the maxTime is bigger than the actual time that took the passenger to arrive at the second boarding station.

4.2.2 Bus-to-bus transfer

To detect a transfer from one bus to another, firstly, a list of the bus stops in the walking area around the next boarding stop is obtained from the nearStop table. After that, a filter is applied to the near stop list to select the ones that the first boarding bus passes through. Thereby, the list of the stops that passengers could arrive at with the initial bus that s/he took is obtained. If there is a bus stop in the zone that fulfills the criteria, the maximum time that could be spent between two consecutive boarding locations is calculated.

Equation 4.3, shows how to calculate the maximum time needed between two smart card transactions in case of aiming to have a transfer. The first term is used to get the travel duration in the first bus, the next term is to calculate the time spent walking from one stop to the next one, and the final term is the average bus waiting time.

$$maxTime = \frac{distance(first\_boarding(lat, lon), candidate\_stop\_location(lat, lon))}{250} + \frac{distance(candidate\_station\_boarding(lat, lon), second\_boarding(lat, lon))}{50} + 10 \tag{4.3}$$

4.2.3 Train-to-bus transfer

When a transfer occurs from rail-based transportation to a bus, the area around the bus stop that the passenger would walk to have the transfer is scanned. A list of the candidate train stations is obtained from the nearStop table. After, it is controlled if any of the candidate stations have a common train line with the first boarding station. If a station that meets these conditions is found, the maximum time is calculated by equation 4.4.

The first term stands for the duration of the first ride made via train, then, the walking

time between the train station and the bus stop, the average bus waiting time, average train waiting time, and the in-station activity time are added.

$$maxTime = \frac{distance(first\_boarding(lat, lon), candidate\_station\_location(lat, lon))}{500} + \frac{distance(candidate\_station\_boarding(lat, lon), second\_boarding(lat, lon))}{50} + 10 + 10 + 5 \tag{4.4}$$

If there are stations located in the buffer zone, but they don't meet the requirement of having a common line with the first boarding station, the case of unobservable metro-to-metro transfer is considered and the maximum time is calculated according to it. By adding 10 extra minutes to the equation 4.4 for the waiting time between the trains, the maximum time is obtained.

If there is no train station in the boundary of the area defined around the second boarding stop, that indicates there is no transfer.

4.2.4 Bus-to-train transfer

When a passenger uses first a bus and then transfers to a train it is easier to detect the transfer compared to the previous scenario. Similar to the previous type of transfer, firstly the buffer zone around the second boarding location, the train station, is scanned to get a list of the bus stops. After that, the bus stops are controlled whether the first boarded bus line passes through these candidate stops. The stop meeting these conditions is considered the alighting stop of the first trip, and the maximum time variable is calculated according to this stop.

Equation 4.5, which consists of only two elements that are the bus travel duration and the walking time between the bus stop and the train station is used for calculating the maximum time needed. In this case, there is no need to add a waiting time or in-station activity time. Because the transaction record timestamp is the time that a passenger enters a bus and starts moving forward and when the passenger arrives at the train station for the next trip, the record is created when s/he enters the station, therefore, the train waiting time is not included.

$$maxTime = \frac{distance(first\_boarding(lat, lon), candidate\_stop\_location(lat, lon))}{250} + \frac{distance(candidate\_station\_boarding(lat, lon), second\_boarding(lat, lon))}{50} \tag{4.5}$$

4.3 Final destination assumption

The traditional trip chaining algorithm assumes that the destination of the last trip of the day is the same location as the first boarding location however, later, this assumption is relaxed by saying somewhere near to the first boarding location instead being exactly the same location. At this point choosing the most suited day-starting hour is crucial. There may be many passengers returning to their houses after midnight, and if the day-starting hour is taken as midnight, 00.00, the destinations of these records cannot be estimated.

In this thesis, the hour that has the lowest transportation activity was taken as virtual midnight, and the algorithms were developed based on this hour. In the Madrid metropolitan area, the most used transportation mode, the metro, stops its operations between 2 AM and 6 AM, and the lowest usage of public transportation is expected to be during this time interval. Figure 4.3 shows the transportation usage for 24 hours, from 1st of February 2020, 00.00 to 2nd of February 2020, 00.00. As it can be seen from the plot, while 11.00, 12.00, and 18.00 are the peaks of the public transportation usage by the senior citizens of the Community of Madrid with around 40.000 trips per hour, 04.00 has the lowest activity with only 51 rides. The number of trips between 00.00 to 01.00 is 1222 which is quite higher than the number of trips at 04.00. Hence virtual midnight of 04.00 is considered the optimal midnight for the trip chaining algorithm.

After having the 24-hours data with virtual midnight, the near stop list for the first trip of the user is obtained. It is checked if there is any station or stops in that list the user can arrive via the last transportation line. If it exists, that stop or station is considered the last destination of the user to arrive at its residential property.

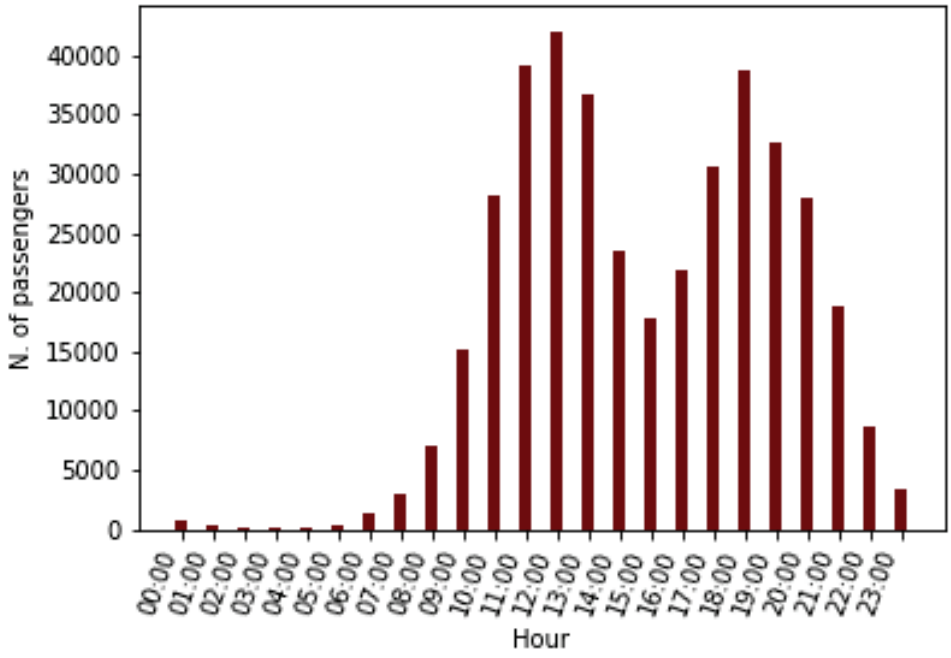

Figure 4.3: Hourly transaction record counts for February 1, 2020.

4.4 OD Matrix Generation

Once the occurrence of the transfer is checked, it is written in a Boolean column. After that, if the trip is the last trip of the user for that day, the final destination assumption is applied to estimate the destination. In case the trip is not the last one, the first trip chaining assumption proposed by Barry et al.[11] is applied. The boarding location of a trip is used to estimate the destination of the previous trip by checking the stops around the boarding location. When a bus stop or a train station is found in that area, it is checked if the passenger can reach it with the previous transportation mode and line. Thereby, the destination of each transaction is controlled. If the destination is estimated, it is saved to a new column. -1 or -2 are written to the destination column respectively when the final destination or intermediate destinations could not be estimated.

After detecting the transfers and estimating the destination for each transaction, the transactions that are detected as transfers are combined into one transaction by taking the boarding value of the first transfer leg and the destination of the last leg of the transfer chain. The destination of the intermediate transfers is not significant because the goal of the passenger is to arrive at the destination of the last transfer leg. Later, the rows that their destination could not be estimated are removed. Thus, the final data frame that contains all the transactions with their origin and destination locations is obtained.

Thereafter, a column that contains the origin-destination stop ID pairs is created. According to the area scale that the OD matrix is desired to be created, the rows of the data frame can be grouped and the number of occurrences of each OD pair can be counted to generate the OD matrix.

The OD Matrix can be created for each bus stop and train station, each district, each postal zone, or each zone. The number of unique stops and stations is around 13 thousand, therefore, it is not appropriate to demonstrate the result of the stop/station pairs in the OD matrix. Instead, the flows between the districts, or zones are more convenient. In the following chapter, the results of the OD matrix that is generated by grouping the stops and stations in the same districts are shown.

5 Evaluation

In this chapter, the results achieved by performing the algorithms described in the previous chapter on the sample dataset mentioned in chapter 3 are presented in 4 sections.

5.1 Virtual Midnight

The algorithm is tested with 3 different time intervals, February 1, 2020, Saturday with actual midnight, the same day with virtual midnight, 04.00, and February 4, Tuesday with virtual midnight, 04.00. Naturally, The amount of data that exists for each time interval is different. Table 5.1 shows the number of transactions executed during these time periods. Moreover, the size of the datasets after the single transactions are removed is shown.

As it can be seen in table 5.1, due to the change of midnight on Feb 1, the number of the total records increase by 489. However, the number of records after excluding the single records increase by 1221. In other words, when the midnight change from 00.00 to 04.00, 732 more records create a trip chain or joins an existing chain. Thereby, the number of the excluded rows drops.

	Feb 1, 00.00 - Feb 1, 23.59	Feb 1, 04.00 - Feb 2, 03.59	Feb 4, 04.00 - Feb 5, 03.59
N. of records	440.597	441.086	709.137
N. of records without single transactions	399.170	400.391	653.709
Ratio of the single transactions	0,094	0,092	0,078

Table 5.1: Number of the transaction records without removing any data and after removing the single transactions with the single transaction ratio.

5.2 Transfer Detection

In figure 5.2, an example of detected transfer record is shown. Firstly, the passenger takes a metro at 16.20 from Manuel Becerra station which is at line 2 and line 6. 24 minutes later, the same passenger uses a bus line 625 from stop 6274 which is located very close to Moncloa metro station. Figure 5.1 shows the boarding locations on a map. As it can be seen Manuel Becerra and Moncloa stations have metro line 6 in common. In the line 6 route, there are 9 stops between these two stations which can be traveled in around 18 minutes. Therefore, the algorithm detects a transfer from inferred alighting station, Moncloa, to the 2nd boarding bus stop location.

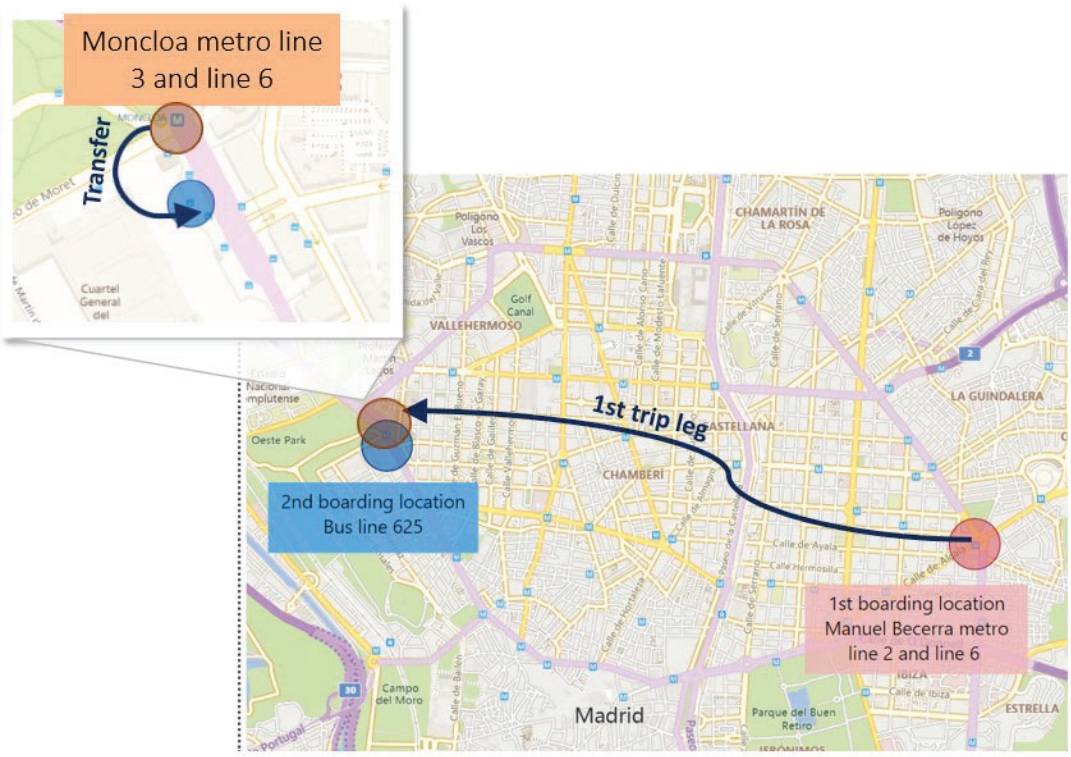

Figure 5.1: A map that with a detected transfer record.

	cardID	date	DPAYPOINT	DENOMINAPARADA	LATITUD	LONGITUD	type	IDLINEA	IDSTOP	transfer
	FFFF9F4806C8E9EE27661C9D62950E81	2020-02-01 16:20:56	02_L2_P2	MANUEL BECERRA	40.42790	-3.66921	0	NaN	156	True
	FFFF9F4806C8E9EE27661C9D62950E81	2020-02-01 16:44:50	26_L625_P1	NaN	40.43428	-3.71929	1	625	6274	False

Figure 5.2: The record of the detected transfer.

	Feb 1, 00.00 - Feb 1, 23.59	Feb 1, 04.00 - Feb 2, 03.59	Feb 4, 04.00 - Feb 5, 03.59
N. of total transfers detected	83.474	83.774	133.610
N. of middle transfers	20.745	20.818	35.203

Table 5.2: Number of transfers detected.

Table 5.2 shows the number of transfers detected for each sample period. The total amount of the observed transfers is given in the first row. When there are more than two consecutive transfers, the transfers between the first and the last leg of the transfer chain are called middle transfers, and their numbers are shown in the second row.

5.3 Destination Estimation

In table 5.3, the number of the trips that their destination is estimated by the proposed algorithm is given. The number in the first row is the estimated destination number by using the dataset that the single transactions and middle transfers in it are removed. The numbers in the next row are obtained from the dataset that all the transactions in the transfer chains, except the first one, are removed in addition to the removed single transaction.

	Feb 1, 00.00 - Feb 1, 23.59	Feb 1, 04.00 - Feb 2, 03.59	Feb 4, 04.00 - Feb 5, 03.59
N. of destination estimated (the single transactions and middle transfers are excluded)	107.686	107.889	168950
N. of destination estimated (all the transfers are excluded)	90.626	90.880	142.345

Table 5.3: Number of the estimated destinations.

While the destination estimation algorithm searches for the candidate destination locations, it also classifies the unsuccessful destination estimations into two groups. The first one is the records that their final destination could not be obtained which means no candidate stop or station found around the first boarding location of the day. The other group is for the intermediate records that there is no alighting location option around

the next boarding area that passengers can reach with the previous transportation mode. The size of the second group is much bigger than the first one 5.4.

	Feb 1, 00.00 - Feb 1, 23.59	Feb 1, 04.00 - Feb 2, 03.59	Feb 4, 04.00 - Feb 5, 03.59
N. of unestimated final destination	61.028	61.313	92.813
N. of other unestimated destination	203.443	203.964	284.959

Table 5.4: Number of the unestimated destinations.

5.4 OD Matrix

The data frame obtained by applying the transfer detection and the destination estimation algorithms to create the OD matrix contains transactions only if their destinations were estimated. In table 5.5, the final estimation ratios are shown. The goal of the trip chaining method is to estimate the destinations of every trip in the dataset after removing the single transactions and combining a transfer chain into one trip. Therefore, the denominator of the ratio equation is the size of the dataset obtained after excluding the single transactions and combining the transfer chains into one row (first row in table 5.5). The dividend is the size of the OD matrix obtained from the proposed algorithm (second row in table 5.5).

	Feb 1, 00.00 - Feb 1, 23,59	Feb 1, 04.00- Feb 2, 03.59	Feb 4, 04.00 - Feb 5, 03.59
N. of records (after excluding single transactions and transfers)	314.579	315.491	520.099
N. of records in the OD matrix	90.626	90.880	142.345
Final estimation ratio	0,28808	0,28805	0,27368

Table 5.5: The final estimation ratios for each time periods.

In the following sections, the OD matrix obtained is examined for an entire day and for the morning, afternoon, and night spans of the day.

5.4.1 24-Hour Period

For the Feb 1, 2020, virtual midnight case, the size of the final OD matrix is 90.880. The origin-destination were grouped by their district codes to observe the interdistrict passenger flow. Map 5.3 was generated by picking the OD pairs that have more than 300 travels during the day from the OD matrix given in table 5.6. The green and the red points respectively represent the origin and the destination of the same district. However, their size demonstrates the amount of the in and outflow. For instance, the Madrid center has more out-flow than in-flow because the size of the green bubble is bigger than the red one. On the other hand, for the rest of the districts the size of the origin and destination bubbles are very similar to each other.

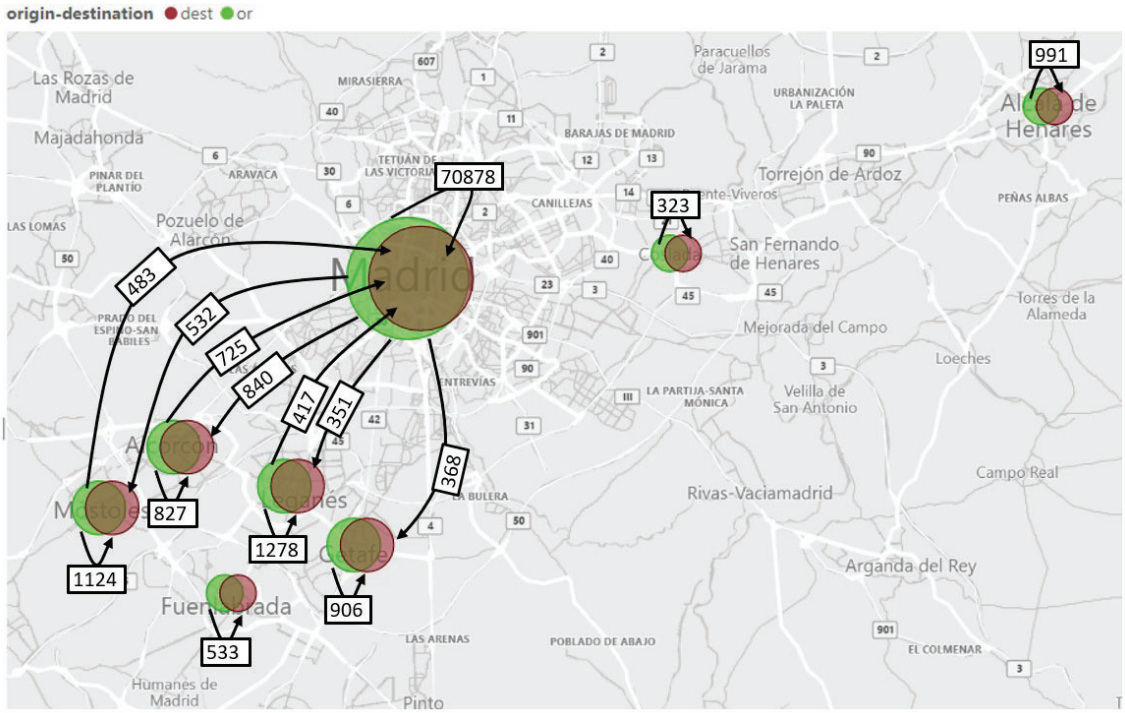

Figure 5.3: Passenger flows that have more than 300 travels during the entire day.

Table 5.6 was generated with 8 districts that have the highest amount of flow. The vertical axis (rows) of the table is the origin of the destinations while the horizontal axis (columns) is the destination locations. Madrid center is the only district that interacts with all the other districts given in the table.

	Madrid 28079	Leganés 28074	Móstoles 28092	Alcalá de Henares 28005	Getafe 28065	Alcorcón 28007	Fuenlabrada 28058	Coslada 28049
Madrid	70878	351	866	-	368	840	-	-
Leganés	417	1248	12	-	55	27	73	2
Móstoles	483	4	1124	-	-	66	27	4
Alcalá de Henares	180	-	-	991	4	-	-	4
Getafe	241	51	6	-	906	12	23	-
Alcorcón	725	22	141	-	8	827	17	-
Fuenlabrada	133	43	55	-	16	7	533	-
Coslada	198	1	4	2	2	2	1	323

Table 5.6: OD Matrix of districts of the Community of Madrid (with their district codes).

5.4.2 Morning, Afternoon and Night Periods

The OD matrix mentioned in the previous section was divided into 3 intervals as shown in the table 5.7. In figure 5.4, 5.5, and 5.6, the districts with the ten highest passenger flow respectively, for the morning, afternoon and night are shown. Additionally, the number of passengers who starts a trip from the Madrid center and who arrives in the Madrid city center is demonstrated by the green and red lines respectively.

As it can be observed from the maps, during the morning period, senior citizens prefer to move forward to the Madrid center, during the afternoon period, in-and-out flow approaches an equilibrium point, and during the night period, the majority move from the Madrid center to other smaller districts.

	Time Interval	Size of the dataset
Morning	04.00-12.00	15.213
Afternoon	12.00-20.00	55.604
Night	20.00-03.59	20.043

Table 5.7: Time intervals and their sizes.

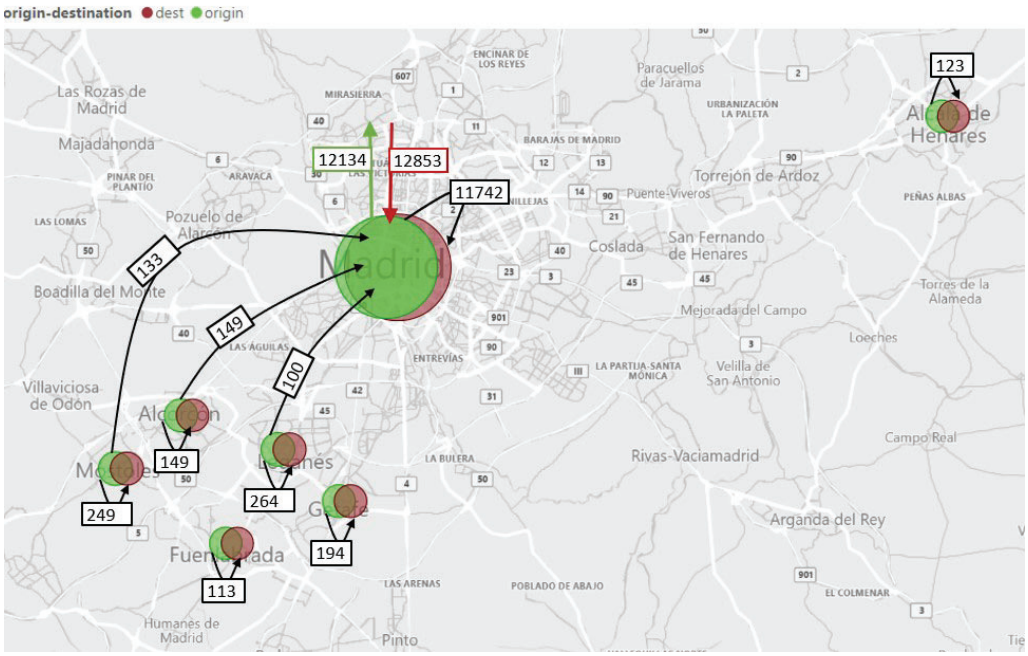

Figure 5.4: Passenger flow in the morning.

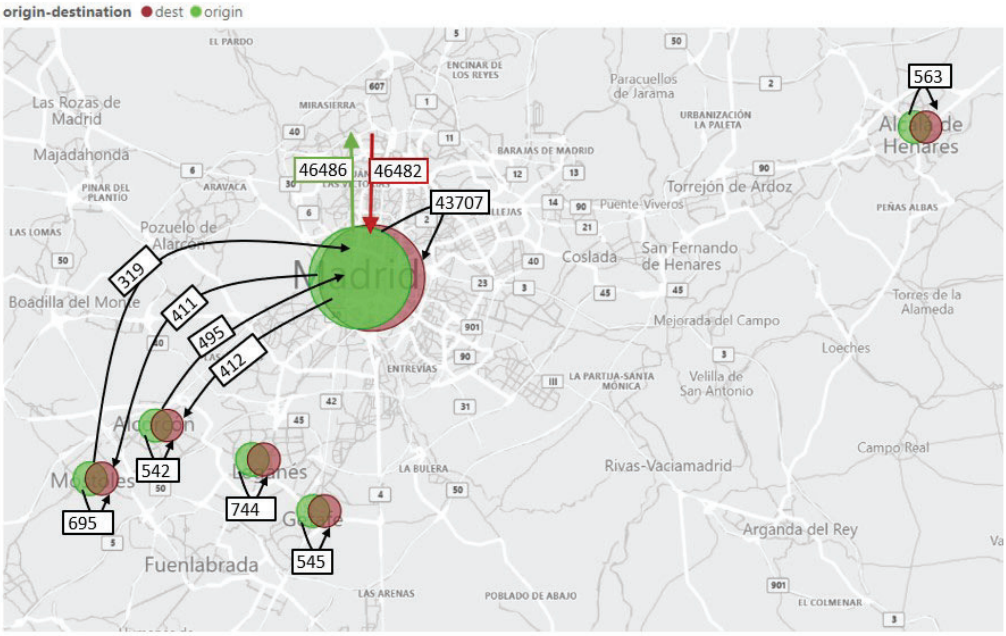

Figure 5.5: Passenger flow in the afternoon.

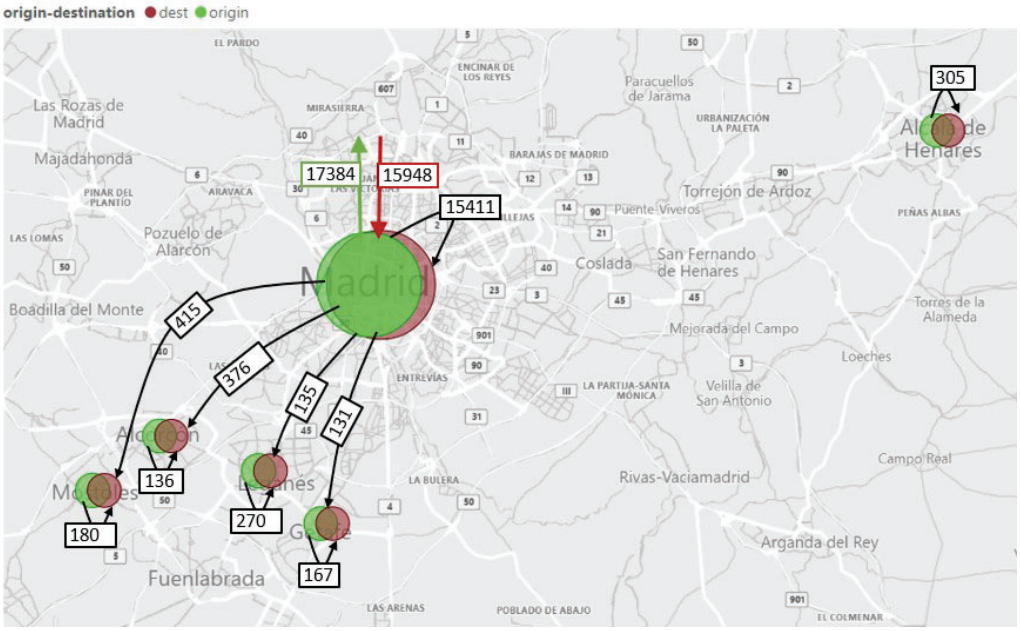

Figure 5.6: Passenger flow in the night.

6 Conclusions

6.1 Conclusions

Three OD matrices for three different time periods were generated by using the trip chaining method developed for the Madrid public transportation system. These matrices preserve numerous knowledge to be mined. The conclusions of this thesis can be enumerated;

- Most of the transactions are performed in the Madrid center or between the Madrid center and other districts.
- Senior citizens primarily prefer to travel inside of a district where their residential property is located more than interdistrict travels. In the case that they leave their districts, the most traveled district is the Madrid center with a vast majority.
- For all the sample data, roughly 1 in 5 trips were detected as transfers. Around 25% of the transfers are the middle transfer leg of a transfer chain (multiple consecutive transfers).
- For 27% of the transactions given into the destination estimation algorithm, the destination location was estimated. The records of the single transaction are not included in this calculation due to the nature of the trip chaining method, estimating the destination of the single transactions is not aimed with the proposed algorithm. This percentage increase to 28.8 when the transfers are combined in one transaction.
- The change of the midnight from 00.00 to the hour that the least amount of transactions performed which is 04.00 decreases the number of single transactions eliminated from the dataset due to the trip chains over the night.

- While the single transaction ratio is 0,092 on Feb 1, Saturday, it drops to 0,078 on Feb 4, Tuesday. That shows the citizens tend to walk or to use a private transportation mode for completing their trip chain at weekends more than the weekdays.
- During the morning, passengers flow forward to the city center, while it is the opposite for the night period. This shows that senior citizens prefer to travel to the Madrid center during the day to perform an activity, later, they travel back to their residential area.
- The districts where the senior citizen activities are observed more are the Madrid center, Móstoles, Alcorcón, Leganes, Fuenlabrada, Getafe, and Alcalá de Henares. These districts are also the most crowded seven districts of the Community of Madrid with a population that around 20% of it is formed by senior citizens older than 65 years old [1]. It shows us that these areas have the majority of the senior citizen's residents, thereby, the direction of the flows can be explained.
- The results show there is no significant difference in the estimated destination percentage between a weekday and a weekend. The ratio of estimation drops slightly for Tuesday, although the increase was foreseen due to the expectation of higher regularity in the travel pattern of weekdays.
- Some transfers detected manually could not detected by the algorithm. The reason is the elimination of some bus stop data due to the missing values on the crucial columns.
- Finally, the estimation ratio, 0.28, shows the difficulty of predicting the destination of a trip by only deploying the trip chaining algorithm, due to the complexity of the passenger travel behaviors.

6.2 Future Works

To improve the performance and the results of the algorithms developed, the following works can be done in the future.

- Performing a deeper analysis of the passenger behaviors in order to select the optimum values for the variables such as maximum walking distance, transfer times, etc.

[1] `http://portalestadistico.com/municipioencifras/?pn=madrid&pc=ZTV21`

- In the bus and metro stops dataset, there are some missing stops. The absence of stops causes to fail to estimate the destination for the trip. Therefore, this dataset should be completed.
- In this thesis, due to the unavailability of the whole passenger dataset, only the records from senior citizens were used. To evaluate the algorithm more accurately, the full dataset is needed to be used.
- Performing a demographic analysis of the areas of the Community of Madrid to evince the characteristic of the neighborhoods. Thereby, a rational explanation of the destinations can be claimed. Moreover, this analysis can be used to validate the estimations.
- The efficiency and the run-time of the implementation can be improved.
- Lastly, by applying the proposed algorithm to a much larger dataset, OD matrices for a longer period of time can be obtained and used for machine learning-based ridership forecasting.

7 References

[1] "Urban transportation systems of 25 global cities," McKinsey & Company, Tech. Rep., 2021. [Online]. Available: https://www.mckinsey.com/~/media/mckinsey/business%5C%20functions/operations/our%5C%20insights/building%5C%20a%5C%20transport%5C%20system%5C%20that%5C%20works%5C%20new%5C%20charts%5C%20five%5C%20insights%5C%20from%5C%20our%5C%2025%5C%20city%5C%20report%5C%20new/elements-of-success-urban-transportation-systems-of-25-global-cities-july-2021.pdf.

[2] "Informe anual 2019," Consorcio Regional de Transportes de Madrid, Tech. Rep., 2019. [Online]. Available: https://www.crtm.es/media/880193/informe_anual.pdf.

[3] J. Ding, M. Yang, Y. Cao, and S. L. Kong, "Dwell time prediction of bus rapid transit using arima-svm hybrid model," in *Sustainable Cities Development and Environment Protection IV*, ser. Applied Mechanics and Materials, vol. 587, Trans Tech Publications Ltd, Aug. 2014, pp. 1993–1997. DOI: 10.4028/www.scientific.net/AMM.587-589.1993.

[4] C. Ding, D. Wang, X. Ma, and H. Li, "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees," *Sustainability*, vol. 8, no. 11, p. 1100, 2016.

[5] L. Heydenrijk-Ottens, V. Degeler, D. Luo, N. van Oort, and J. van Lint, "Supervised learning: Predicting passenger load in public transport," in *CASPT Conference on Advanced Systems in Public Transport and TransitData*, 2018, pp. 30–32.

[6] K. Pasini, "Forecast and anomaly detection on time series with dynamic context. application to the mining of transit ridership data.," Ph.D. dissertation, Université gustave eiffel, 2021.

[7] F. Toqué, E. Côme, L. Oukhellou, and M. Trépanier, "Short-term multi-step ahead forecasting of railway passenger flows during special events with machine learning methods," in *CASPT 2018, Conference on Advanced Systems in Public Transport and TransitData 2018*, 2018, 15p.

[8] C. Ding, J. Duan, Y. Zhang, X. Wu, and G. Yu, "Using an arima-garch modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, pp. 1054–1064, Mar. 2018. DOI: 10.1109/TITS.2017.2711046.

[9] A. Lacki, "Analysis and characterization of the public transport mobility of senior citizens," M.S. thesis, Universidad Politecnica de Madrid, 2019.

[10] M. P. Beldarrain, "Analisis y prediccion individual del comportamineto de los usuarios con abono tercera edad en el transporte publico de la comunidad de madrid," M.S. thesis, Universidad Politecnica de Madrid, 2021.

[11] J. J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda, "Origin and destination estimation in new york city with automated fare system data," *Transportation Research Record*, vol. 1817, no. 1, pp. 183–187, 2002. DOI: 10.3141/1817-24. eprint: https://doi.org/10.3141/1817-24. [Online]. Available: https://doi.org/10.3141/1817-24.

[12] J. Zhao, "The planning and analysis implications of automated data collection systems: Rail transit od matrix inference and path choice modeling examples," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[13] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 1–14, 2007. DOI: 10.1080/15472450601122256. eprint: https://doi.org/10.1080/15472450601122256. [Online]. Available: https://doi.org/10.1080/15472450601122256.

[14] A. Alsger, B. Assemi, M. Mesbah, and L. Ferreira, "Validating and improving public transport origin–destination estimation algorithm using smart card fare data," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 490–506, 2016, ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2016.05.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X16300353.

[15] A. A. Nunes, T. Galvao Dias, and J. Falcao e Cunha, "Passenger journey destination estimation from automated fare collection system data using spatial validation," *Trans. Intell. Transport. Syst.*, vol. 17, no. 1, pp. 133–142, Jan. 2016. DOI: 10.1109/TITS.2015.2464335. [Online]. Available: https://doi.org/10.1109/TITS.2015.2464335.

[16] J. B. Gordon, H. N. Koutsopoulos, N. H. M. Wilson, and J. P. Attanucci, "Automated inference of linked transit journeys in london using fare-transaction and vehicle location data," *Transportation Research Record*, vol. 2343, no. 1, pp. 17–24, 2013. DOI: 10.3141/2343-03. eprint: https://doi.org/10.3141/2343-03. [Online]. Available: https://doi.org/10.3141/2343-03.

[17] A. A. Alsger, M. Mesbah, L. Ferreira, and H. Safi, "Use of smart card fare data to estimate public transport origin–destination matrix," *Transportation Research Record*, vol. 2535, no. 1, pp. 88–96, 2015. DOI: 10.3141/2535-10. eprint: https://doi.org/10.3141/2535-10. [Online]. Available: https://doi.org/10.3141/2535-10.

[18] N. Nassir, A. Khani, S. G. Lee, H. Noh, and M. Hickman, "Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system," *Transportation Research Record*, vol. 2263, no. 1, pp. 140–150, 2011. DOI: 10.3141/2263-16. eprint: https://doi.org/10.3141/2263-16. [Online]. Available: https://doi.org/10.3141/2263-16.

[19] F. Yan, C. Yang, and S. V. Ukkusuri, "Alighting stop determination using two-step algorithms in bus transit systems," *Transportmetrica A Transport Science*, vol. 15, no. 2, pp. 1522–1542, 2019, ISSN: 2324-9935. DOI: https://doi.org/10.1080/23249935.2019.1615578. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2324993522002901.

[20] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012, ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2012.01.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X12000095.

[21] W. Wang, "Review on hybrid flow shop scheduling," in *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, vol. 4, 2011, pp. 7–10. DOI: 10.1109/ICM.2011.219.

[22] M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 70–79, 2014, ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2014.03.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X14000801.

[23] A. Cui, "Bus passenger origin-destination matrix estimation using automated data collection systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[24] E. Hussain, A. Bhaskar, and E. Chung, "Transit od matrix estimation using smartcard data: Recent developments and future research challenges," *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 103 044, 2021, ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103044. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X21000759.

[25] T. Li, D. Sun, J. Peng, and K. Yang, "Smart card data mining of public transport destination: A literature review," *Information*, vol. 9, p. 18, Jan. 2018. DOI: 10.3390/info9010018.

[26] J. J. Barry, R. Freimer, and H. Slavin, "Use of entry-only automatic fare collection data to estimate linked transit trips in new york city," *Transportation Research Record*, vol. 2112, no. 1, pp. 53–61, 2009. DOI: 10.3141/2112-07. eprint: https://doi.org/10.3141/2112-07. [Online]. Available: https://doi.org/10.3141/2112-07.

[27] A. Ali, J. Kim, and S. Lee, "Travel behavior analysis using smart card data," *KSCE Journal of Civil Engineering*, vol. 20, no. 4, pp. 1532–1539, 2016.

[28] M. Bagchi and P. R. White, "The potential of public transport smart card data," *Transport Policy*, vol. 12, no. 5, pp. 464–474, 2005.

[29] J. Liu and X. Zhou, "Observability quantification of public transportation systems with heterogeneous data sources: An information-space projection approach based on discretized space-time network flow models," *Transportation Research Part B: Methodological*, vol. 128, pp. 302–323, 2019.

[30] M. MOSALLANEJAD, S. SOMENAHALLI, A. VIJ, and D. MILLS, "An approach to distinguish destination from the alighting stop based on fare data," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 13, pp. 1348–1360, 2019.

[31] L. He, N. Nassir, M. Trépanier, and M. Hickman, *Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems.* CIRRELT, 2015, vol. 52.

[32] M. Hofmann and M. O'Mahony, "Transfer journey identification and analyses from electronic fare collection data," in *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005.*, IEEE, 2005, pp. 34–39.

[33] P. Kumar, A. Khani, and Q. He, "A robust method for estimating transit passenger trajectories using automated data," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 731–747, 2018.

[34] K. K. A. Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transportation research record*, vol. 2063, no. 1, pp. 63–72, 2008.

[35] C. Seaborn, J. Attanucci, and N. H. M. Wilson, "Analyzing multimodal public transport journeys in london with smart card fare payment data," *Transportation Research Record*, vol. 2121, no. 1, pp. 55–62, 2009. DOI: 10.3141/2121-06. eprint: https://doi.org/10.3141/2121-06. [Online]. Available: https://doi.org/10.3141/2121-06.

[36] M. Yap, O. Cats, N. van Oort, and S. Hoogendoorn, "A robust transfer inference algorithm for public transport journeys during disruptions," *Transportation research procedia*, vol. 27, pp. 1042–1049, 2017.

[37] D. Huang, J. Yu, S. Shen, Z. Li, L. Zhao, and C. Gong, "A method for bus od matrix estimation using multisource data," *Journal of Advanced Transportation*, vol. 2020, 2020.

[38] H. Dou, H. Liu, and X. Yang, "Od matrix estimation method of public transportation flow based on passenger boarding and alighting," *Computer and Communications*, vol. 25, no. 135, p. 79, 2007.

[39] X. Zhou, X. Yang, and X. Wu, "Origin-destination matrix estimation method of public transportaion flow based on data from bus integrated-circuit cards," *Journal of Tongji University. Natural Science*, vol. 40, no. 7, pp. 1027–1030, 2012.

[40] Y. Wanbo, W. Hao, Y. Xiaofei, X. Chuangchuang, and J. Dongxue, "Od matrix inference for urban public transportation trip based on gps and ic card data," *Journal of Chongqing Jiaotong University (Natural Science)*, vol. 34, no. 3, p. 117, 2015.

[41] Z. Mengmeng, G. Yajuan, and M. Yujiao, "A probability model of transit od distribution based on the allure of bus station," *Journal of Transport Information and Safety*, no. 3, pp. 57–61, 2014.

[42] Y. Jie and X. Yang, "Estimation a transit route od matrix using on/off data: An application of modified bp artificial neural network," *Syst. Eng*, vol. 24, pp. 89–92, 2006.

[43] J. Jung and K. Sohn, "Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data," *IET Intelligent Transport Systems*, vol. 11, no. 6, pp. 334–339, 2017.

[44] L. Zhang, "Study on the method of constructing bus stops od matrix based on ic card data," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, IEEE, 2007, pp. 3147–3150.

[45] J. M. Farzin, "Constructing an automated bus origin–destination matrix using farecard and global positioning system data in sao paulo, brazil," *Transportation research record*, vol. 2072, no. 1, pp. 30–37, 2008.

[46] J. Zhao, A. Rahbee, and N. H. Wilson, "Estimating a rail passenger trip origin-destination matrix using automatic data collection systems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 376–387, 2007.

[47] D. Li, Y. Lin, X. Zhao, H. Song, and N. Zou, "Estimating a transit passenger trip origin-destination matrix using automatic fare collection system," in *International Conference on Database Systems for Advanced Applications*, Springer, 2011, pp. 502–513.

[48] O. Egu and P. Bonnel, "How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? an empirical investigation in lyon," *Transportation Research Part A: Policy and Practice*, vol. 138, pp. 267–282, 2020.

[49] B. Assemi, A. Alsger, M. Moghaddam, M. Hickman, and M. Mesbah, "Improving alighting stop inference accuracy in the trip chaining method using neural networks," *Public Transport*, vol. 12, no. 1, pp. 89–121, 2020.

A Appendix - GitHub Repository

Pre-requisites installation

The libraries required in order to run the notebooks in the Repository[1];

- Python 3.9 Official Site[2]
- Pandas Official Site[3]
- NumPy Official Site[4]
- Numba Official Site[5]
- Matplotlib Official Site[6]

There are 3 notebook files in the given repository;

- **near_stop.ipynb**
  This file has to be run to obtain the near stops dataset mentioned in the thesis. This dataset is required to run the next notebooks.
- **destination_transfer.ipynb**
  This file has a pipeline that firstly reads the datasets and preprocesses them. Later defines the transfer detection functions, and creates a data frame with all the records for the chosen time interval with the transfer value, and

[1] https://github.com/dogacengiz/TFM
[2] https://www.python.org/downloads/
[3] https://pandas.pydata.org/
[4] https://numpy.org/
[5] https://numba.pydata.org/
[6] https://matplotlib.org/

estimated destination locations. Finally, this data frame is saved as CSV.

- **OD_matrix.ipynb**
  This notebook takes the CSV file created from the previous step, and processes it to obtain the desired OD matrix. The example OD matrix is created by grouping the locations based on the district codes as explained before.

Additionally, the manually generates lines.csv which stores the metro-line pairs is shared in the repository.