Forecasting spatial and temporal variations in OD-pairs

-

A case study from São Paulo

Forecasting spatial and temporal variations in OD-pairs

A case study from São Paulo

Version: Final

Author: Jan M. Engels
(s1175718)

February 25, 2019

Supervisors:
Prof. Dr. Ing. K.T. Geurs (Karst)
Prof. Dr. M.A. Giannotti (Mariana)
Dr. K. Gkiotsalitis (Konstantinos)

UNIVERSITY OF TWENTE.

ESCOLA POLITÉCNICA DA
UNIVERSIDADE DE SÃO PAULO

Faculty of Engineering Technology
Centre for Transport Studies (VVR)
University of Twente
The Netherlands

Acknowledgements

After spending more than 25 weeks in Brazil this is the final piece of work of my time as a student. This report presents the work that has been executed within the research project that is required to earn the title 'Master of science'. Before presenting the work, results and conclusions I would like to take the chance to thank some people in person which made this research project possible.

First of all I would like to thank my supervisors Prof. Dr. Ing. Karst Geurs, Prof Dr. Mariana Giannotti and dr. Konstantinos Gkiotsalitis, their feedback was always helpful and helped me to achieve the set objectives and finalize the research project. During my stay in São Paulo the help from colleagues at the 'LABGEO' was also a valuable addition to my projection, thank you for your help.

Also I would also like to thank the 'Van Eesteren-Fluck & Van Lohuizen Stichting'. The foundation grants subsidies to projects in the field of urban planning and landscape architecture in the spirit of Van Eesteren and Van Lohuizen. The foundation financially supported my work which made going to Brazil for me possible.

Abstract

In the need for more sustainable urban transport the recent years have seen the deployment of so called public bike sharing systems (PBSS). These are systems that provide bicycles to customers on an as needed basis. Customers can pick up the bicycles at fixed locations and return them to the same or a different station. Since May 2012 such a system is also available in São Paulo, Brazil.

As the number of systems worldwide increased they also gained the attention of the academic society. They investigated which factors influence the usage, for what purposes people use the system and how the bike sharing system can be integrated with the existing public transport opportunities of a city. Most of the researches focussed on the surrounding factors of a station and how they can influence the number of arriving and/or departing trips. From a system-operator point of view it is more interesting to know how the bicycles move within the system. Until now there is only a limited amount of literature available that investigates the movements between different stations within a public bike sharing system. This is the main reason why this research, next to stations, looked at the number of trips between OD-pairs. The factors that have been identified for other PBSS will be tested for the São Paulo case. It is unknown if the factors that influence the number of arrivals/departures at station level also apply for OD-pairs. Therefore the objective of this research is to develop a model that can forecast spatial and temporal variations in the number of trips between OD-pairs in a PBSS.

The factors that have been identified from the literature are; availability of bicycle infrastructure around a station, land use around the station, proximity to other transport modes, environmental aspects (slopes and weather) and population characteristics. One of the challenges of this research is to identify if these factors also apply to OD-pairs. For this research the historical data of rental processes from 2017 has been used to identify the number of trips. For the surrounding factors different data sources from the municipality and the LABGEO have been used.

The first analysis was to identify the temporal patterns of the bicycle usage. The outcomes show that two clear patterns can be identified, one for days during the week and one for weekends. For week days a morning and an evening rush-hour peak can be identified. For the weekends no peak could be identified. The rental duration also differs, during the weekend the customers tend to use the bicycles for longer periods than during the week. Looking at the most frequent used stations it can be seen that from Monday till Friday most trips start and end in proximity to areas with high commercial activities. During the weekend the bicycles are more frequently picked up at stations that are close to recreational areas, like parks.

The next step of the analysis was to identify if the factors pointed out in the literature also apply to the trip attraction and generation in São Paulo. The outcomes of the linear regression models show that not all factors apply. If the variables are significant only a small portion of the variation is explained by the model, all models have low R-squared values.

Based on findings from the trip attraction and generation it has been tested if these factors are also suitable to describe spatial variations. To control if the factors are also applicable to OD-pairs linear regression models have been set up with the number of trips per OD-pair as dependent variable. These linear regression models have been separated by population density and by trips during the week and weekend. The strongest correlation was obtained between the number of trips and the average daily temperature. The models show that if the average temperature increase more people use the system. Looking at the travel times between the stations it has been discovered that in denser populated areas people tend to make shorter trips. The other variables that may apply to trip attraction and generation at the station level do not necessarily apply to OD-pairs.

This research modelled spatial and temporal variations in the number of trips. As the results show the historical average of the number of trips per OD-pair per time stamp is a good approximation. However to detect trends and to improve the forecasts the difference between the predicted and observed number of trips from the previous time stamp should be taken into account. Taking this information into account resulted in halving the size of the average residual. The average residuals were further decreased by specifying the model for warmer and colder temperatures. The outcomes of the model validation showed that the historical average rental time is not a good predictor for the current rental duration.

Summarizing the findings from the research it can be said that the factors that have been identified for stations by other researches do not all apply for the case of São Paulo. Also these variables are not very suited to describe spatial and temporal variations in OD-pairs. One hypothesis is that other factors like personal preferences and socio-economic characteristics also influence the number of trips. These factors have not been studied in this research. However clear spatial and temporal patterns in the number of trips of OD-pairs can be obtained from the 2017 data. There are clear differences in the usage between week days and weekends. To forecast these variations in the number of trips the historical average is a suitable approach. In combination with information from previous time stamps and the weather the model can very accurately forecast the number of trips.

Reflecting on the research it has to be mentioned that it was assumed that during all moments there were enough bicycles and docking points available at each station. Unfortunately it was not possible to get data about the amount of available bikes or docking points. For further research one should take the number of trips per bike per day into account and investigate the attitude towards cycling in São Paulo.

Management summary

The recent years have seen many developments in the field of sustainable transport. One of these trends is the deployment and development of public bike sharing systems in various forms. These bike sharing systems provide the customers with bicycles on as needed basis. Since 2012 such a system is available in São Paulo, Brazil. The system, that is analysed in this research, allows the customers to pick up the bicycles from fixed stations. The system that is in operation in São Paulo consists of 261 of these fixed stations.

The first system that can be considered a bike sharing system was deployed in 1965 in Amsterdam, the Netherlands. Since then the systems and especially the bikes used within the system have developed. They improved from simple unlocked white painted bikes to sturdy bikes with online locking mechanism and GPS-tracker. Since the middle of the 1990's researchers have started investigating and analysing these systems. Since this period the number of systems that operate in different cities across the world also increased.

As the number of systems worldwide increased they gained the attention of the academic society. There are many different researches that analyse various aspects of the bike sharing systems. From a city planner point of view researchers identified strategies on how to integrate these system into the existing public transport system. Another aspect that has been studied with respect to PBSS is for what trip purposes the bicycles are used. For the operator of the systems it is very interesting how the bicycles move within the system and between stations. The topic that has been studied the most extensively, are the factors that influence the trip attraction and generation of each station.

The factors that influence the number of arriving and departing trips per station and that have been identified by the literature are; surrounding available bicycle infrastructure, land use around the station, proximity to other transport modes, environmental aspects (slopes and weather) and population characteristics. With respect to the available infrastructure many researches have shown that the quality and presence can positively influence the number of trips (e.g. [Cleland and Walton, 2004]). When it comes to the land use, a more divers land use around a station can increase the number of arriving and departing trips per station (e.g. [McBain and Caulfield, 2017]). The distance to connecting public transport plays an important role for the number of trips per station. If the distance to other public transport modes from a station decreases the number of trips increases [Raux et al., 2017]. The city characteristics like topography and size have been identified by the research of [Cleland and Walton, 2004] to have an influence on the number of trips. The distance from home or work to a station is another factor that influences the number of trips per station [Wang and Lindsey, 2019].

Next to identifying the factors that influence the number of trips per station, researchers also tried to forecast the number of arriving/ departing trips or the number of available bikes per station for given moments in time. For these forecasts different approaches like Markov-chains or ARIMA-models have been used. Until now there is limited number of researches that focussed on forecasting the number of trips between stations, so called OD-pairs and therefore forecasting spatial variations in the number of trips. It is unknown if the factors at the station level, that have been identified by other works, apply to the case of São Paulo.

Based on the gaps in the literature the following research objective was defined: *Develop a model that can forecast spatial and temporal variations in the number of trips between OD-pairs in a PBSS.*

First of all the factors that have been identified for other systems will be tested for the case of São Paulo. It is unknown if the factors that influence the number of arrivals/departures at station level also apply for OD-pairs. Therefore it will be tested if the factors that apply to stations in the PBSS of São Paulo are also suitable to describe spatial variations. The unique thing of this research is that the research goes beyond just identifying the factors that influence the usage at the station level. This approach is possible due to the amount and quality of the available data.

This research will use different data-sources to identify the factors. For the surrounding characteristics of each station information from the municipality and the LABGEO will be used. The individual trip data, the most important information for this research, was provided by the Brazilian centre for analysis and planning (CEBRAP). For this research the historical data of rental processes from 2012 until 2017 has been used to identify the number of trips. This researches includes 261 stations and 438.862 individual trips.

The first part of the analysis focussed on identifying the temporal patterns of the bicycle usage. From the graphs two clear different patterns can be observed. The first pattern includes all trips that occurred between Monday until Friday. During the weekend a different pattern was observed. The historical data from days during the week clearly shows two peaks in the usage of the system. The first peak is during the morning rush-hour and the second during the evening rush-hour. In the weekend no peaks were obtained, the number of trips slowly increases, stays the same during the day and then decreases again in the evening hours. Other differences that were obtained between the week and weekend are differences in the rental duration. During the weekend customers use the bicycles for longer periods than during the week. Also during the weekend bicycles are picked up and returned at the same station.

Looking at the most frequent used stations it can be seen that from Monday till Friday most trips start and end in proximity to areas with high commercial activities. During the week the most frequent used stations are the one located close to the metro station 'Faria Lima'. This area is attractive to cycle because of a high station density and a lot of bicycle infrastructure is present in this area. The most used stations during the weekend are stations that are located at the entrance to 'Parque Ibirapuera', the biggest park in São Paulo. When one looks at the number of trips per OD-pair it can be seen that the OD-pairs with the highest number of trips are located in the area around 'Faria Lima'. The analysis showed that in the morning the bicycles are picked up close to the metro station and in the evening returned to a station close to the metro station.

The second part of the analysis focussed on the factors that influence the trip attraction and generation at the station level. The factors that have been previously mentioned are used in a linear regression model as independent variables. The dependent variables are the number of trips arriving or departing from each station. Looking at the infrastructure it can be said that the model is significant but the explanatory power is very low. However the coefficients are in line with the expectations from the literature, more available bicycle infrastructure in proximity to the station results in more trips. The relation between number of trips and land use was not discovered for the case of São Paulo. Looking at the relation between the distance to public transport and the number of trips something strange was observed. When the distance to a public transport mode increases the number of trips also increases. This is the opposite of what one expect from the literature. The job density around the stations also has a significant impact on the number of arriving and departing trips. However the R-square value of the model is about 12%.

After identifying the factors that influence the trip attraction and generation, the factors were used to forecast spatial variations. The unit of analysis for this part was the number of trips between OD-pairs and not the number of arriving and departing trips per station. The challenge herein was the factors around the origin and destination of a trip had to be considered. The outcomes of the linear regression models all have very low R-squared values.

The strongest correlation was obtained between the number of trips and the average daily temperature. The models show that if the average temperature increase more people use the system, this is in line with the expectations from the literature review. Another important finding from the analysis of the factors that influence the number of trips between OD-pairs is that the travel times between the stations in denser populated areas are shorter. This means people travel more frequent between station that are located close to each other.

As previously mentioned the objective of this research is to develop a model that can forecast spatial and temporal variations in the number of trips between OD-pairs. Therefore after identifying the factors that influence the trip attraction and generation as-well as the spatial variation, the historical data was used to set up the model. The data was first grouped based on the OD-pair per trip and then the average number of trips per time-stamp has been calculated. This was done for each of the 8.327 individually OD-pairs to make a database. For the forecast the model inputs are the code of the origin, the code of the destination and the desired time-stamp of the forecast.

The model includes a part that checks for trends in the data and then adjusts the forecast. For the identification of the trend the information from the previous time-stamp is most suitable. The model has been validated in two different ways. Firstly it was controlled if the historical average per time-stamp is already a good estimate. The average residual of the model was -0,4 this means per time stamp there were 0,4 more bikes estimated than actually travelled between the OD-pair. By incorporating the information from the previous time stamp the average residual was reduced by about 50%. As the analysis shows the average daily temperature influences the number of trips. Therefore the model has been separated again for warmer and colder days. By doing so the average residual was further decreased to about -0,25.

The model also used the average of the historical rental duration to give an estimation on when the bicycles will arrive at the destination. However during the validation it was discovered that this is not a suitable approach.

Summarizing the findings from the research it can be said that the factors that have been identified for stations by other researches do not all apply for the case of São Paulo. Also these variables are not necessarily suited to describe spatial variations in OD-pairs. One hypothesis is that other factors like personal preferences and socio-economic characteristics also influence the number of trips. These factors have not been studied in this research. However clear spatial and temporal patterns in the number of trips of OD-pairs can be obtained from the 2017 data. There are clear differences in the usage between week days and weekends. To forecast these variations in the number of trips the historical average is a suitable approach. In combination with information from previous time stamps and the weather it can very accurately predict the number of trips.

Looking at the results of the research it must be assumed that other factors that have not been studied influence the usage of the system. One of these factors can be personal preferences like attitude towards cycling or feeling of safety during the trip. Also the operational aspect should not be neglected. Unfortunately there was no information available about the number of working bikes per station per moment in time or how many docking points were available to return the bikes. These variables can strongly influence the number of trips between OD-pairs. For further research one should focus on getting more insights into these aspects for better understanding and describing the behaviour of the PBSS in São Paulo.

Contents

1 Introduction **12**

1.1 General . . . 12

1.2 Case of São Paulo . . . 12

1.3 General information about bike sharing systems . . . 14

1.4 Readers guide . . . 15

2 Literature review **16**

2.1 Trip generation and attraction . . . 16

2.2 Demand estimation and trip purposes . . . 17

2.3 Spatial distribution of trips . . . 17

2.4 Spatial and temporal variation in the usage . . . 18

2.5 Research on how and why people use bicycles in São Paulo . . . 19

2.6 Methods to forecast the usage . . . 19

2.7 Clustering techniques . . . 21

2.8 Summary literature review and contribution of the research . . . 22

3 Research objectives and questions **23**

3.1 Research objectives . . . 23

3.2 Research questions . . . 23

4 Data used for this research **25**

4.1 Data types . . . 25

4.2 Data preparation . . . 26

4.3 Quality of the data . . . 30

4.4 Summary data used for this research . . . 32

5 Analysis **33**

5.1 Spatial and temporal patterns . . . 33

5.2 Factors that attract or generate trips per station . . . 38

5.3 Factors influencing spatial variations . . . 42

5.4 Summary analysis . . . 50

6 Forecasting the number of trips between OD-pairs **51**

6.1 Preparing the data . . . 51

6.2 Calculating the necessary data . . . 51

6.3 Model input and data selection . . . 53

6.4 Forecasting the number of trips . . . 53
6.5 Validation number of trips . . . 54
6.6 Validation rental duration . . . 58
6.7 Summary forecasting the number of trips between OD-pairs . . . 59

7 Conclusions **60**

7.1 Identify the factors that influence the usage of PBSS in São Paulo . . . 60
7.2 Analyse the spatial and temporal variations in the number of trips between OD-pairs of the PBSS in São Paulo . . . 60
7.3 General conclusions . . . 61

8 Discussion **62**

Appendix A Outputs **64**

A.1 Additional graphs and maps . . . 64
A.2 Distance decay functions . . . 70
A.3 Likelihood of change in job accessibility . . . 71
A.4 Parameters of regression models (Weather) . . . 72
A.5 Factor analysis . . . 72

Bibliography **74**

List of Figures

1 Location of the system 13
2 Full extend of the system 14
3 Main components of the PBSS 14

4 Snapshot of data 27
5 Land use around the system [Secretaria Municipal de Financas 2016] 28
6 Schema for population density calculation 29
7 Population density per area [Census (2010)] 30

9 Travel and rental time per time stamp 34
10 Most frequent used stations 35
11 Land use around metro station 'Faria Lima' 36
12 Trips per OD-pair during the morning peak (Week) 37
13 Trips per OD-pair during the evening peak (Week) 37
14 Types of bicycle paths 38
15 Relation between infrastructure and arrivals/departures 39
16 Relation between travel time and number of trips 43
17 Distance decay function (Week) 43
18 Distance decay function and fits 44
19 Relation between available infrastructure and the number of trips 45
20 Regression models with temperature as independent variable 46
21 Exponential decay function of the job accessibility (5 minutes service area) 47
22 Exponential decay function of the job accessibility (10 minutes service area) 47

23 Visualization of the model 52
24 Residuals (Week) - regular pattern only 55
25 Residuals (Weekend) - regular pattern only 55
26 Residuals (Week) - Using information from previous steps 56
27 Residuals (Weekend) - Using information from previous steps 56
28 Residuals (Week) - Separated by temperature 57
29 Residuals (Weekend) - Separated by temperature 57
30 Residuals of rental duration (Week) 58
31 Residuals of rental duration (Weekend) 58

32 Relation between distance to public transport and arrivals/departures 64
33 Regression models for distance to public transport (Week) 65

34 Regression models for distance to public transport (Weekend) 66
35 Regression model for maximum slope between origin and destination 66
36 Regression model for change in land use between origin and destination 67
37 Change in job accessibility between origin and destination 67
38 Cumulative density plot of the number of trips per OD-pair 68
39 Topography of São Paulo 68
40 Bicycle infrastructure in São Paulo 69
41 Metro and CPTM lines of São Paulo 70

List of Tables

1 Overview of bicycle sharing generations . . . 15

2 Results from the research of [Fecchio, 2018] . . . 19

3 Data related to the bicycle trips . . . 25
4 Data related to the environment . . . 25
5 Number of stations per year . . . 26
6 Number of trips per month/day . . . 31

7 Stations with the most arriving and departing trips . . . 35
8 Amount of available bicycle infrastructure . . . 39
9 Regression results infrastructure and trip attraction/generation per station . . . 40
10 Regression results land use mix and trip attraction/generation per station . . . 40
11 Regression results distance to public transport and trip attraction/generation per station 41
12 Regression results population and job density and trip attraction/generation per station . 42
13 Regression results temperature and number of trips . . . 46
14 Extracted factors (Week) . . . 48
15 Extracted factors (Weekend) . . . 48
16 Comparison of the regression results . . . 50

17 95% confidence interval of residuals (Only historical data) . . . 55
18 95% confidence interval of residuals (With trend) . . . 56
19 95% confidence interval of residuals (With temperature) . . . 58

20 Fitted parameters for the exponential function . . . 70
21 Fitted parameters for the Richard's function . . . 71
22 RMSE per fit (Week & Weekend) . . . 71
23 Fitted parameters for the Richard's function (5 minutes service area) . . . 71
24 Fitted parameters for the Richard's function (10 minutes service area) . . . 72
25 RMSE for Richard's function . . . 72
26 Regression model weather (Estimated coefficients) . . . 72
27 Load factors for variables (Weekend) . . . 72
28 Load factors for variables (Weekend) . . . 73
29 T-test results (Week) . . . 73
30 T-test results (Weekend) . . . 74

1 Introduction

1.1 General

The recent years have seen the adoption of many new transport modes like ride hailing, self driving cars, bike sharing systems and very recent the implementation of electric scooters. Researchers all over the world are looking for new possibilities and modes of transportation as the need for transportation increases while the space, especially in urban areas, is limited. These new technologies and concepts, like for example MaaS (Mobility as a Service), have the potential to change the way how society moves within a city. This research will look at one of these 'new transport modes', bike sharing systems, that might help to solve transportation problems in urban areas.

Various types of bike sharing systems have been introduced in different cities across the world. These so called public bike sharing systems (PBSS) have a wide number of advantages. First of all they are easy to use for citizens and secondly they have the potential to reduce greenhouse gas emissions in cities, these are only two of the possible benefits. In the development over the years these PBSS have become essential parts of the public transport systems for cities. However these systems have various challenges that researchers try to cope with.

When one looks at the different stages of implementing and operating a bike sharing system three major challenges, that are related to the different phases, can be identified. The first is estimating the scale of operation. This means how many bikes are required and which area the system will serve. The second challenge is to forecast the number of trips that are made within the system. This challenge is closely related to the problem of rebalancing the system. The operators want to ensure enough bikes across the service area to satisfy the demand of the customers. To ensure that the demand of bikes is met the operators have to reposition the bikes. Repositioning means picking up the bikes where the demand is low and transporting them to areas where the demand is high. This repositioning of the bikes is resource consuming for the operator, therefore a lot of effort is put into improving this process. This brings us to the third challenge. Knowing where the bicycles will be in the future. The number of trips between stations depends on different factors such as time and spatial characteristics of the station. These two components make the prediction of the number of trips between stations very complicated. There are other minor challenges like e.g. assuring sufficient maintenance of the bicycles. This research will investigate the factors that influence the numbers of trips within a PBSS. The challenge herein lies in the fact that the number of trips depends on time and space. The research-problem, that this research will try to answer, can be summarized to 'How can spatio-temporal variations in the number of trips between OD-pairs be foreseen?'.

1.2 Case of São Paulo

This research focusses on one specific PBSS. The bike-sharing system is in operation in the city of São Paulo, Brazil. This section will give a short description of the city and in which context the system is operated. On the top right hand side of figure 1 a map of the country Brazil can be seen. The area that is marked in green is the province of São Paulo. A bigger map of the province São Paulo can be seen in the left top corner of figure 1. The area that is marked red in the map of the province is the city of São Paulo. In the metropolitan area of São Paulo live about 21 million people, in the urban centre about 12 million. The average elevation is around 799m above sea level. The third map in figure 1 shows the location of the bicycle sharing system within the city of São Paulo. It can be seen that the system is located in the north of the city. This area is the most dense populated area and where most commercial activities take place. Figure 2 zooms in to the area where the stations are located. This figure shows all the stations that are used in this research.

The system in São Paulo is called 'Bike Sampa' and it was deployed in May 2012. The system started with 100 station spread across the centre of the city. Until 2014 the system was extended to 163 stations. From 2014 the 2015 the number of stations increased the most to 261 stations. In 2016 and 2017 this number slightly decreased to 252. So it can be seen that since the launch in 2012 the system expanded. Since 2018 the system is operated by 'Tembici'. The two main elements of the system are the bicycles and the stations, see figure 3. These two images show the bikes and stations that were newly introduced by 'Tembici' in 2018. Before that the bikes and stations had different designs but still their recognizable orange colour.

In general the bicycles are designed in a robust way to ensure that they can stay in operation for a long time. The number of bikes that can be locked at a station varies and depends mostly on the demand. The stations are mostly located within, what can be considered as the centre of São Paulo. In north-south direction the stations are located between the river 'Tiete' and the neighbourhood 'Campo Belo'. In the west the 'Pinheiros' river forms a natural barrier to the system, in the east the system operates till the neighbourhood 'Vila Matilde'. The stations are located between 719m and 827m above sea level.

All the other necessary information that is required to better understand the system and the number of trips between stations will be given in one of the following chapters.

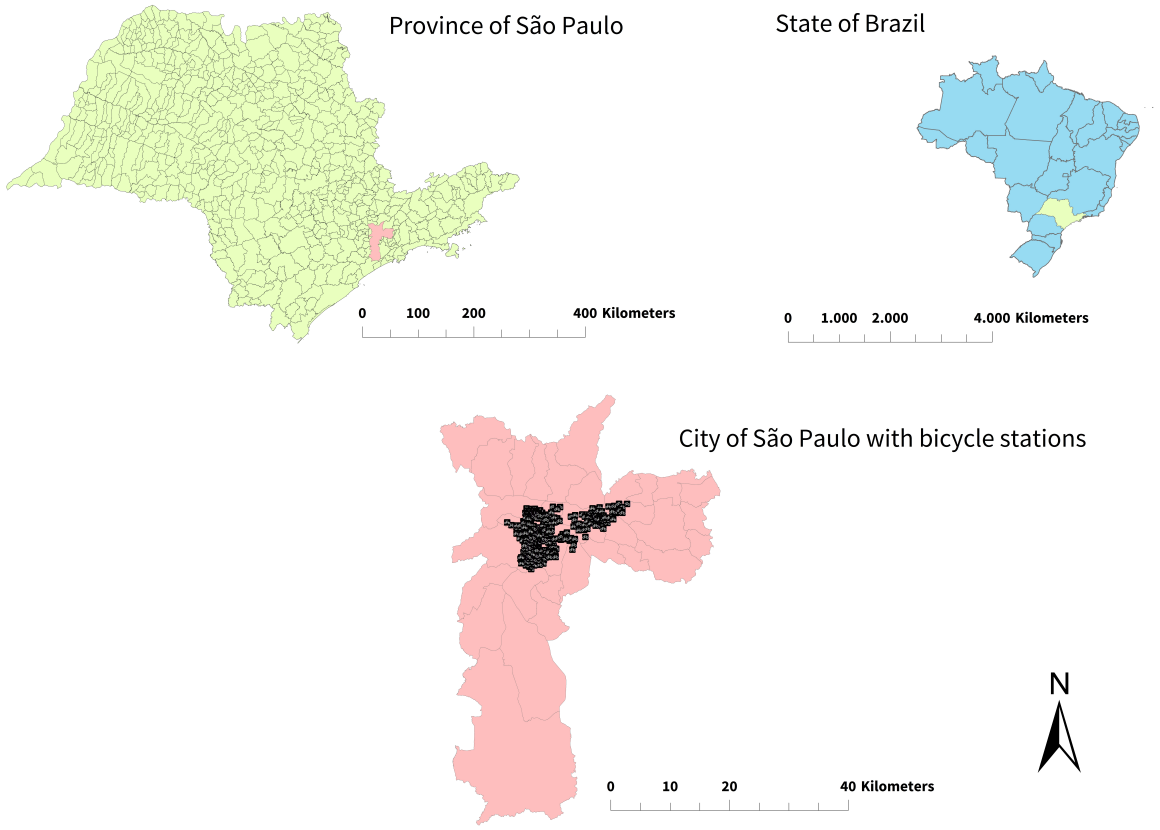


Figure 1: Location of the system

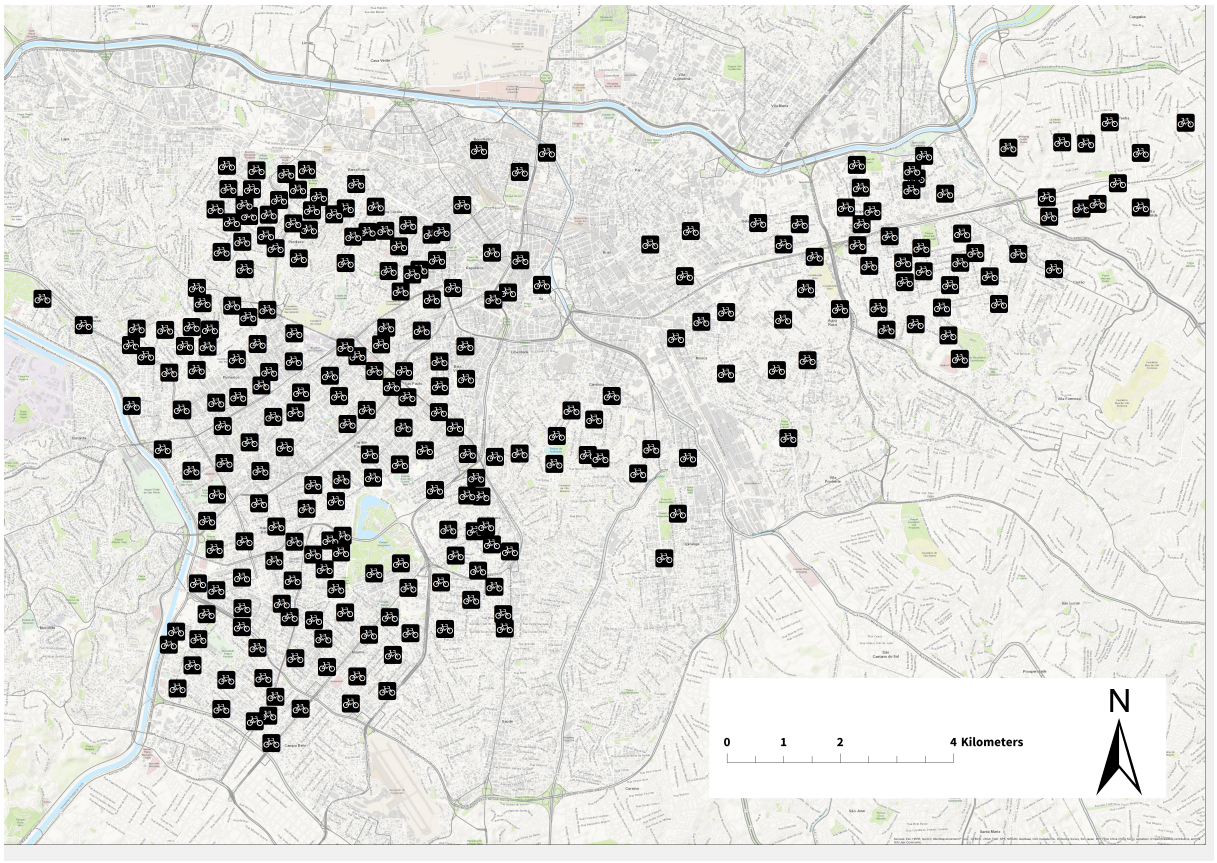


Figure 2: Full extend of the system

Figure 3: Main components of the PBSS

1.3 General information about bike sharing systems

'Bicycle sharing systems (BSS) are the first new form of public transportation in over a hundred years to be widely adopted' [Médard de Chardon, 2016]. According to [Demaio, 2009] one can distinguish between three generations of bike sharing systems. The first was the so called 'Witte Fietsen' (White Bikes) in Amsterdam. This system was launched in 1965 and contained simple bikes which were painted white and then provided to the public. The second generation started in 1991 in Denmark. In the second generation the bikes have been improved to be more appropriate to the utilitarian use. Similar to the first generation, the second generation had the problem of theft due to the anonymity of the users. The first third generation bike sharing system was started in 1996 at the Portsmouth University in England. For this system the users needed a magnetic stripe card to rent a bike. This identification of users is the main characteristic that distinguishes the third generation from the previous two. Over the years a

fourth generation was developed. The system that is currently in operation in São Paulo is one of the fourth generation. An overview of the different generations has been made by [Shaheen et al., 2010] and can be found in table 1.

The latest development are so called 'free floating bicycle sharing systems'. They can be seen as the fifth generation. Due to new technologies they do not require stations. The technology to unlock and lock the bikes are integrated within the bikes. The bikes can be theoretically picked up and returned everywhere. Using a mobile application the customers can look where the bicycles are located and unlock the bicycles. In São Paulo such a system is also, next to the system from 'Tembici', in operation. The fifth generation has not been mentioned by [Shaheen et al., 2010] and has been manually added to the table.

Generation	**Components**	**Characteristics**
1	1. Bicycles	1. Distinct bicycles 2. Bicycles located haphazardly throughout an area 3. Bicycles unlocked 4. No charge for use
2	1. Bicycles 2. Docking stations	1. Distinct bicycles 2. Bicycles located at specific docking stations 3. Bicycles with locks
3	1. Bicycles 2. Docking stations 3. Kiosk or user interface technology	1. Bicycles are distinct 2. Bicycles are located at specific docking stations 3. Bicycle have locks 4. Smart technology is used for bicycle check-in and checkout 5. Theft deterrents 6. Programs are paid for as a membership
4	1. Bicycles 2. Docking stations 3. Kiosks-user interface 4. Bicycle distribution system	1. Distinct bicycles 2. Program may include electric bicycles 3. Specific docking stations that are more efficient 4. Improved locking mechanism to deter bicycle theft 5. Touch screen kiosks-user interface 6. Bicycle redistribution system 7. Linked to public transit smart-card
5	1. Bicycles with technology to unlock the bicycles 2. Bicycle distribution	1. Distinct bicycles with GPS 2. Program may include electric bicycles 3. Improved locking mechanism to deter bicycle theft 4. Bicycle redistribution system

Table 1: Overview of bicycle sharing generations

1.4 Readers guide

The remaining of the report is structured as follows. First a literature review executed to identify the work that has already been done in this field and to determine where there are gaps within the current knowledge. Based on the identified gaps the research objectives and research questions can be formulated. For each objective the methodology that is going to be used to achieve the objectives will be described in the same chapter. Because this research is a data driven research the next chapter will describe the data that is used. After the description of the available data, the analysis of which factors influence the number of trips between OD-pairs will be executed. This chapter is then followed by the description of the model that can forecast the temporal variations of the number of trips between OD-pairs. The final two chapters will present the conclusion and the discussion of the research. All these chapters will have the same structure, first of all a general brief introduction will be given followed by the content. The chapters will be concluded with brief summary.

2 Literature review

As previous mentioned this chapter will put the research in perspective of the current work in the field of PBSS. Over the past years as the PBSS evolved, the scientific community has put a lot of effort in understanding the behaviour of the system and which factors influence the usage of the systems. The following sections will describe the various factors that are relevant to this research. This includes analysing the literature about factors that influence the usage, trip attraction and generation, demand estimation and forecasting trips between OD-pairs.

2.1 Trip generation and attraction

Before it will be explored how people might use the PBSS it is important to understand why people decide to use the systems. There are various factors that influence the fact if people want or do not ant to use the PBSS.

The first factor that has been identified is the *available infrastructure* [Cleland and Walton, 2004], [McLoughlin et al., 2012], [McBain and Caulfield, 2017], [Nielsen and Skov-Petersen, 2018]. There are two types of infrastructure that have been identified. Firstly the road infrastructure like bicycle paths. The other type is related to the bike sharing system itself. In the case of the system the term infrastructure refers to bikes and stations. These researches identified that the quality and the availability of the two types of infrastructure are an important factor why people might cycle/ use the system.

The next factor is the *land use* around the origin and destination of a bicycle trip. A diverse land use around the stations promote the usage of the PBSS [McBain and Caulfield, 2017], [Nielsen and Skov-Petersen, 2018] and [Zhang et al., 2017]. A research that focussed especially on the factors that influence the number of bicycle trips in Brazil identified that people who cycle, consider a diverse land-use around their destination as an important factor [Freitas and Maciel, 2017a].

The third factor are *city specific characteristics* like size of the city, topography and weather. In Brazil [de Souza et al., 2017] identified that the size of the city can have an influence on the bicycle usage. In this relation the topography (e.g. do people have to cycle steep gradients) plays a major role [Cleland and Walton, 2004], [McLoughlin et al., 2012]. A map of the topography of São Paulo and the stations can be found in the appendix. The weather plays an important role when people are deciding if they are going to cycle or not. The influence of high temperatures and/or strong rain on the bicycle usage has been identified by [McLoughlin et al., 2012].

The fourth factor is the access to a PBSS and the *connection to other transport modes*. A close proximity of the stations to other transport modes can increase the usage of the PBSS-station, that has been identified by various studies from different systems [Raux et al., 2017], [Zhao and Li, 2017] and [Fuller et al., 2011]. The quality of the connecting transport mode can also influence the usage, 'Faster and higher quality types of public transport, such as train and intercity buses, attract substantially more bike-and-ride users than slower and lower quality types of public transport, such as local buses or trams' [Martens, 2004]. A research by [González et al., 2016] states that that 'the model shows that in the system studied, destination choice is strongly influenced by Metro stop locations, indicating that a combined bicycle-Metro mode generates a strong synergy'.

Another factor is the *access to the system*. The accessibility to the system is normally measured in the distance to the nearest station. A research that as conducted on the 'Nice-Ride' system in Minneapolis and St. Paul, USA, discovered that 'The proximity to bike share stations from residence has a significant impact on bike share use by members' [Wang and Lindsey, 2019]. Extending the system across a whole city and close proximity to a PBSS can positively effect the usage [Raux et al., 2017]. In Beijing the 'travel distances between home and transition zones was found to be the most important factor influencing people's decisions to cycle or not' [Zhao and Li, 2017]. In the case of Montreal, Canada, it was discovered that the people living within 250m from the PBSS-stations use the system more often than people living more far away [Fuller et al., 2011]. Another factor that influences the access to the system is that some systems require a credit card to register or a deposit to use the system. This is especially a problem for low income groups [Gauthier et al., 2013].

Summarizing the findings from the literature, the factors that influence the usage of PBSS are:

- Infrastructure (Availability and quality)
- Land use (Type and diversity)
- City characteristics (Size, topography and climate)
- Connection to other transport modes (Quality of mode and distance to mode)
- Access to the system (Distance and registration)

2.2 Demand estimation and trip purposes

The researches that focus on the topic of demand estimation concentrate mostly on the time before a system is deployed in a city. From a policy-maker and operator point of view it is interesting to identify where and how high the demand in the future will be. The results of these kind of researches can help to determine the size and location of the future system/ stations. Most of these approaches determine the demand based on socio-economic characteristics of areas and the demand of other transport modes. For the demand estimation of a PBSS in Philadelphia, USA, [Krykewycz et al., 2010] put a grid across the city. Based on the characteristics of each cell they then determined the travel demand. A similar approach has been taken by [Ji et al., 2014] to estimate the demand for a electric bike sharing system at a university campus in the UK.

The PBSS can generally speaking be used for two trip purposes; recreational and work related. According to [Zhang et al., 2018] the system are mostly used for work/ school related purposes. In Brazil the bicycle forms an important transport mode to commute to work [Freitas and Maciel, 2017b]. Next to the two very general purposes researchers also identified in which situations or for what trips customers use the systems. The first situation is to make the whole trip from origin to destination using a PBSS. In Zhuzhou, China, a significant portion of the users used the PBSS to complete their whole journey by bike [Zhang et al., 2018]. Zhang et al. also identified that users of the PBSS shifted from using their own private bike to using a bike from the PBSS. Another situation where bike sharing systems are used is to access or egress other transport modes. To a slow public transport (bus, metro) people are willing to cycle between 2 or 3 km while for fast public transport (long distance train) people are willing to cycle up to 5 km [Martens, 2004]. When the PBSS is used in combination with public transport, in the case of Rio de Janeiro, Brazil, it has been observed that the quality of the public transport needs to be high [de Souza et al., 2017]. However Zhang et al. identified that PBSS can not always work as a feeder mode for public transport, because the role of PBSS in the transport system depends on the individual situation. With respect to the egress trips it was identified that: 'Egress trips have a smaller average distance that could be ex- tended by increasing the availability of bicycles' [Shelat et al., 2018].

2.3 Spatial distribution of trips

For all transport planners across the world information about the origin and destination of trips is essential for designing their transport networks. 'Origin–destination (OD) flows are a fundamental object of interest in the study of urban transportation networks' [Menon et al., 2015]. Therefore this section focusses on why it is important, which factors influence the flows and how these can be estimated. The number of trips between different OD-pairs is caused by a 'spatial separation of economic and social activities' [Immers and Stada, 1998]. The estimation of the travel demand between OD-pairs is an essential part of the first two steps of the classic four stage model that is used to forecast transport demand in the future.

According to Menon et al. 'the OD matrix is a valuable tool for understanding and forecasting usage patterns of a network' [Menon et al., 2015]. The importance of these OD-matrices has also been pointed out by [Hui et al., 2010], [Lu et al., 2013] and [Peterson, 2007]. Especially interesting to researchers is the estimation of time-dependent OD demand matrices because the separation between social and economic activities can vary throughout time. However it is very difficult to obtain these time-dependent OD

demand [Ma and Qian, 2018]. Peterson adds that these time-dependent OD-matrices can be helpful for strategic and operational purposes.

After identifying why it is important to analyse the flows between origin and destination it is also interesting to identify which factors influence the amount of flow. The research of [Immers and Stada, 1998] identified the following factors that influence the production of a zone, that means the number of trips departing from a zone:

- Household characteristics (e.g. Income and composition)
- Zone characteristics (e.g. socio-economic characteristics)
- Accessibility

For the number of trips arriving at a zone, this is called the attraction, the following factors were identified:

- Number of employees
- Land-use
- Accessibility

According to a research by [Tsekeris and Tsekeris, 2015] the generation from and attraction to a zone are influenced by 'socio-economic, demographic and land use characteristics of each zone'. These factors are very general and can be applied to all forms of transport. As previous mentioned PBSS can be seen as part of the public transport system, therefore it is also interesting to see which factors influence the demand of public transport. According to [Polat, 2012] these factors can be grouped into two categories:

- Structural
  - Cost
  - Quality of service
  - Travel time and distance
  - Alternative transport modes
  - Purpose of travel
- External
  - Demographic
  - Economic and social factors
  - City built environments
  - Policy approaches

These factors that influence the demand are very general and it is unknown to what extend they are applicable to the demand between of OD-pairs of a PBSS.

2.4 Spatial and temporal variation in the usage

This research is especially interested in the spatial and temporal variation in the number of trips between OD-pairs. Therefore the literature that identified these patterns will be explored in more detail. Different researches found that the location of a station influences the temporal pattern of the trips [Froehlich et al., 2009] [Vogel and Mattfeld, 2011] [Etienne and Latifa, 2014] [McBain and Caulfield, 2018]. As already mentioned, weather can influence if people cycle or not in general. Researches showed that the weather can also influence the temporal pattern [Kim, 2018] [Corcoran et al., 2014]. A research that compared the usage in eight different U.S. cities discovered that the spatial and temporal patterns differ between persons that use the systems for commuting purposes and people who use it as tourists [Kou and Cai, 2019].

Until now the focus of researches was on spatial and temporal variations of the number of trips arriving and departing from stations and not on variations in the number of trips between OD-pairs.

2.5 Research on how and why people use bicycles in São Paulo

The previous sections focussed only on the relation between physical measurable units and the number of trips. To identify if personal preferences have also an influence on the number of trips the research of [Fecchio, 2018] can be useful. His research identified the factors that influence the bicycle usage in accessing the metro in São Paulo. The results can help to identify why people maybe use PBSS in São Paulo or not. However when looking at and interpreting the results one has to keep in mind that the research took place in 2018 and that the system is different from the system in 2017. Since 2018 there are fewer stations compared to 2017 and the area in which the stations are located is smaller.

The research included a questionnaire on why people use the bicycles or shared bicycles to access the metro. The question/statements that are relevant for this research are:

- I would use the bike on rainy days
- I would use the bicycle if I have slopes on the route
- Having a bike path on the way would make me use the bike to access the metro
- Being able to use a shared bike, like Itau's bikes, on my way would make me use my bike

The participants of the research were able to give a value between zero (strongly disagree) to ten (strongly agree) for each of the statements. The results can be seen in table 2, from this table it can be seen that the presence of bike paths and a PBSS would make people use a bicycle to access the metro. Also it can be seen that when it rains many people do not want to use a bike to get to the metro. The research also included questions of how respondents go from their origin to metro and then from the metro to their final destination. The results show that only 3% use a PBSS to get from the origin to the metro and that 7% use a PBSS to get to their final destination.

Statements	**Score**
I would use the bike on rainy days	1,99
I would use the bicycle if I have slopes on the route	4,15
Having a bike path on the way would make me use the bike to access the metro	8,04
Being able to use a shared bike, like Itau's bikes, on my way would make me use my bike	7,38

Table 2: Results from the research of [Fecchio, 2018]

The respondents who came by PBSS to the metro were also asked how often they use the system to get to the metro. About 80% of them answered that they use the system four or more times a week. Of the trips that respondents made with a PBSS to get to the metro, 70% cycled only on bicycle paths.

2.6 Methods to forecast the usage

The previous sections focussed on the factors that influence the bike sharing systems. This section will investigate the methods that can be used to forecast the usage of the bike sharing system. Forecasting and discovering the spatial & temporal variations is the main focus of this research.

The modelling approaches can be divided into two groups; the first are models that focus on station-level predictions and the second type are models that try to forecast the movements system wide. Both approaches will be studied in this section. In section 2.8 the literature with respect to forecasting will be evaluated and the 'modelling approach' for our research will be determined.

2.6.1 Station based forecast

The researches that are described in this sub-section focussed on individual independent stations. However it might be that information from neighbouring station will be used to improve the forecasting method.

The researches that focussed on the station level mostly try to forecast the number of available bikes. There are two general approaches that have been taken; the first is using time-series techniques and the other is using probabilistic approaches.

The approaches that use probabilistic techniques intend to determine the 'probability of checking in and checking out between two stations for different moments in time' [Yang et al., 2016]. Yang et al. included also spatio-temporal and meteorology factors in their model to create a dynamic network. Another approach that has taken by [Gast et al., 2015] is to assume that the stations can be modelled using a Markovian model. Gast et al. assumes that the customers arrive according to a Poisson process. With their model Gast et al. can forecast the state of each station. Modelling the rental process/ availability of bikes like a Markov Chain has also been done by [Feng et al., 2017]. They used a 'Population Continuous Time Markov Chain' (PCTMC) to predict the number of available bikes at each station. According to Feng et al. their model outperforms spatio-temporal 'Auto Regressive Integrate Moving Average' (ARIMA) models. A similar approach has been taken by [Guenther and Bradley, 2013]. They used PCTMC's and a modelling technique from the bio-chemical reactions field to determine the number of available bikes. A very simple and straight-forward technique to model the number of available bikes at a station is using linear regression. This approach has been chosen by [Rixey, 2013] and included a lot of characteristics that define each station.

Another interesting approach has been taken by [Borgnat et al., 2011]. They used a technique that is commonly used in signal processing and data analysis. According to them stations have two features - nonstationarity and cyclostationarity. For their approach they modelled both features separately and combined them later on.

Modelling the number of bikes at the station as a time-series is another research area. Before actually modelling the number of bikes/trips many researchers have clustered or organized the stations in some way. A number of works created hierarchical clusters based on rental pattern while others created Voronoi diagrams. According to [Yoon et al., 2012] the 'Voronoi'-approach showed better results than the KNN-approach (K-nearest-neighbour). Using time-series analysis [Vogel and Mattfeld, 2011] were able to model temporal rental patterns. To model these time-series most works used 'Auto Regressive Moving Average' (ARMA) or ARIMA models [Yoon et al., 2012] [Kaltenbrunner et al., 2010]. Kaltenbrunner and Yoon also discovered that including information from neighbouring stations can improve the performance of the model.

2.6.2 System wide forecasts

In contrast to forecasting independent stations this section will go further and describe the literature that focussed on forecasting the distribution of trips. The approaches that have been taken in this field are more diverse, varying from deep learning algorithms and neural networks till linear regression.

To predict the usage of the PBSS in New York City, USA, [Singhvi et al., 2015] aggregated the data to neighbourhood level and then ran regression for every OD-pair. Special about this approach was that they have taken the taxi usage to model the PBSS usage. They used the taxi data because the found 'it particularly useful for predicting pairwise demand, and propose a neighbourhood approach in analysing flows between stations' [Singhvi et al., 2015]. Another approach is to use a 'Structural Equation Model' (SEM) that takes into account the characteristics of the origin and destination to forecast the number of trips. This technique has been used by [Rixey and Ranaiefar, 2016]. To classify their stations Rixey and Ranaiefar also used hierarchical clustering methods. A machine learning model to predict the traffic in a PBSS has been applied by [Li et al., 2015]. The machine learning model they used is a 'Gradient Boosting Regression Tree' (GBRT). To determine the traffic between regions Li et al. applied also a clustering method.

Researchers have tried to apply techniques from other scientific areas to develop a model that can forecast the system-wide usage. In this case [Côme et al., 2014] used a 'Latent Dirichlet Allocation' (LDA) model. This approach is also used for text categorization. The previously mentioned works have focussed on PBSS that use stations where customers can pick up and return their bikes. Xu et al. also modelled the traffic between OD-pairs for a free floating bike sharing system in Nanjing City, China, [Xu et al., 2018]. To do so they used a 'long short-term memory neural network'. Xu et al. used in their model the weather, air quality and land use to train the model. According to the research the neural network performs better, in forecasting the number of trips between stations, than commonly used statistical models.

2.7 Clustering techniques

As mentioned in the previous section many works used various clustering techniques to group OD-pairs or stations. This section will therefore take a closer look at different clustering techniques. In their approach to model bikeshare station usage [Hyland et al., 2018] used both the hierarchical and fuzzy clustering technique to group the stations. They clustered the stations based on the 'type of trips' they attract. The same idea of clustering the stations based on their type of trip was also applied by [Vogel and Mattfeld, 2011], but instead of looking at which type stations attract, they looked at which type stations generate. Next to the hierarchical clustering technique Vogel et al. also used density based clustering techniques. They identified five different types of trips stations can generate:

- Commuter I - High morning trip generation and high evening trip attraction.
- Commuter II - High morning trip attraction and high evening trip generation
- Leisure - High evening and night-time activity
- Tourist - Very high daytime attraction and generation of trips
- Average - trip attraction and generation of trips

Two other works that used hierarchical clustering for grouping the stations are the researches by [Froehlich et al., 2009] and [Lathia et al., 2012]. They, however, clustered the stations based on their occupancy. The work of [Lathia et al., 2012] created next to six clusters, three higher-level clusters in which the stations have been organized. The three higher-level clusters are daytime origins, daytime destinations, and evenly distributed pickup and drop-off.

An approach that has been taken by [Zhang et al., 2018] is to cluster the OD-pairs based on their trip duration rather than stations. This approach might be very useful for our research. A different approach than hierarchical clustering has been taken by [Wang et al., 2016]. They used Moran's I, a statistical tool that determines the spatial autocorrelation between stations, to cluster the stations.

From the available literature it becomes clear that most of the works used hierarchical clustering techniques to group/cluster the data, but they cluster based on different station characteristics or OD-pair characteristics. These clustering techniques can be useful to discover similarities between OD-pairs or individual stations.

2.8 Summary literature review and contribution of the research

Looking at the wide range of available literature with respect to PBSS it can be seen that researchers have put a lot of effort in better understanding the behaviour of this new form of transportation. The factors that influence the trip generation and attraction of stations have been identified for many systems, however they might vary from situation to situation slightly. The benefits of using a PBSS or cycling in general have been extensively studied.

The knowledge about forecasting the number of trips per station is quite extensive while the literature about forecasting the number of trips between OD-pairs is limited. Probabilistic models are frequently used to predict the number of trips per station. To the best of the author's knowledge, there is no work that used probabilistic models to predict the number of trips between OD-pairs.

For the case of São Paulo it is unknown if the factors that attract or generate trips also apply. This research will therefore first focus on identifying the factors that generate and attract trips in the PBSS of São Paulo. With this knowledge this research will create a model that can can forecast spatial and temporal variations in the number of trips between OD-pairs in the case of São Paulo. Going beyond only identifying the factors that influence the usage of the system is the unique characteristic of this research and its main contribution. This combination of identifying the attraction/generation and distribution of trips is unique and is possible due to the available amount of data.

3 Research objectives and questions

This chapter will present the research objectives and research questions. These elements are the core of the research. The answers to the research questions are required to achieve the set objectives. Both elements determine the direction and scope of the research.

3.1 Research objectives

The objectives should be used to determine the direction of the research and identify what lies within the scope of the research. In general this research will be a practice-oriented research. Within the field of practice-oriented research there are a number of types of researches that can be conducted. There are two types of research that are applicable to the previous introduced general problem. The first is diagnostic research, in diagnostic research the background and history of a problem is identified. One has to keep in mind that without effective diagnosis the prognosis will not be effective. The second research type that can be applied is the design-oriented research. In design-oriented research the main objective is to produce new knowledge. Both research types fit within the problem context of this research project. The diagnostic part of the research project will identify if there is really a problem and what the influencing factors are. The design-oriented part will contribute to develop knowledge that can help to solve the problems that have been identified.

Based on the literature review and the gaps that have been identified it is now possible to determine the main research objectives this project tries to achieve. The main research objectives is:

Develop a model that can forecast spatial and temporal variations in the number of trips between OD-pairs in a PBSS.

To achieve this main objective three sub-objectives are set. These sub-objectives are:

1. Identify the factors that influence the usage of PBSS in São Paulo.
2. Analyse the spatial and temporal variations in the number of trips between OD-pairs of the PBSS in São Paulo.

3.2 Research questions

After establishing the research objectives the next step is to formulate the research questions. The two main requirements for research questions are efficiency and steering capacity. Efficiency in the context of research questions means that the answers to the questions should develop knowledge that can help to achieve the set objectives. The research questions should also have steering capacities. This means the questions should describe which steps are necessary in the course of the research. Generally speaking it can be said, efficiency reflects back and the steering function looks ahead to the research activities.

The research questions are separated per research objective. Also a description of the methodological aspects is given per objective.

3.2.1 Identify the factors that influence the usage of PBSS in São Paulo

The research questions that will help to achieve this objective are:

1. What are the influencing factors of bicycle usage of a PBSS in São Paulo?
    - To what extend do the factors that have been identified in the literature review apply to the PBSS in São Paulo?

Methodology: As mentioned in chapter 2 there are various factors that influence the trip attraction and generation. First the necessary data with respect to the factors has to be collected. The next step is then to identify if there are relations between the factors and the number of trips per station. For the number of trips the historical data will be used. Each factor will be analysed separately to identify the impact of each factor. Also different clustering techniques will be applied to observe results for certain groups or regions of stations. The relevant analyses will be presented in section 5.2.

3.2.2 Analyse the spatial and temporal variations in the number of trips between OD-pairs of the PBSS in São Paulo.

The research questions that will help to achieve this objective are:

1. What bicycle usage patterns can be distinguished based on historical PBSS data from the 'Tembici' system?
    - What are the most used origins and destinations in the system of 'Tembici'?
    - What are the travel times between the different stations?
    - What are the characteristics of the different routes?

2. How do travel distances vary across different moments during the day?

Methodology: Again the historical data is required to identify the patterns, most frequent stations and travel times. Using the information from the analysis that identified the factors that influence the trip attraction and generation it will be investigated if these factors are also usable to describe spatial variations in the number of trips.

Using the historical data a probabilistic model will be developed to forecast the number of trips between given origins and destinations for a given moment in time. This model will also include an estimation when the bicycles will arrive at the destination station. The necessary analyses to answer the mentioned research question will be presented in section 5.1.

The answers to all the research questions will help to better understand spatio-temporal variations in the number of trips between OD-pairs in a PBSS.

4 Data used for this research

For this research various types and sources of data are required. This chapter will describe these types and sources. First a general description of the data types and sources will be given, followed by a quality check of the data.

4.1 Data types

Each of the following tables describes one type of data that has been used in this research, they will also give a short description of the data and the source. The abbreviation "CEBRAP" in table 3 stands for 'Centro Brasileiro de Análise e Planejamento' the Brazilian centre for analysis and planning.

Name	Description	Source
Bicycle flow	Origin and destination of trips, including time of departure and arrival (2012 -2017)	Bike Sampa/ Tembici usage data (provided by CEBRAP)
Bicycle stations	Name and location of Bike Sampa / Tembici rental stations (2012-2017)	CEBRAP

Table 3: Data related to the bicycle trips

The 'Bicycle flow' data and the location of the stations are necessary to calculate how many trips departed and/or arrived at each station and how many trips have been made between each OD-pair. These two data sources and sets are the backbone of this research

Name	Description	Source	Spatial unit
Land use	Fifteen different categories of land use that describe the most dominant land use per area	Secretaria Municipal de Financas 2016	-
Topography	Level curves and elevation of Sao Paulo	Secretaria Municipal de Urbanismo e Licenciamento (2004)	m
OD-zones	Zones of origin and destination in São Paulo	METRO (2007)	ha
Road/ Bike network	Roads in Sao Paulo that can be used by car, bus and bicycles	Centro de Estudos da Metropole (CEM) (2007)	km
Accessibility	Job accessibility by bike and ride (BnR) and public transport (PuT)	Laboratorio de Geoprocessamento (2018)	jobs/area
Public transport	Location of public transport stops	SMT/METRO (2018) & SPtrans (2005)	-
Population density	Number of inhabitants per area	Census (2010)	inh/ha

Table 4: Data related to the environment

All the elements that are presented table 4 are necessary to test if the elements, that have been identified in section 2.1, attract or generate trips at the station level in the case of São Paulo. For the land use no spatial unit is defined because the available data shows only the predominant land use for an area. The public transport stations also do not have a spatial unit because the data shows the exact location of each stop. The abbreviation 'inh' in the row 'Population density' stands for inhabitants.

4.2 Data preparation

The previous types of data have all a unique way in which they have been stored. Even the data regarding the bicycles trips has been stored in different forms through the years. The following sections will describe the steps that have been taken to prepare the data for the analysis.

4.2.1 Preparing the station data

As previous mentioned the data with respect to the names and locations of the stations is available from 2012 until 2017. The number of stations per year can be found in table 5. The number of stations that will be used for analysis, 261, is higher than the number of stations in 2017, 252, because stations that might have been closed in 2017 but were open for example in 2012 are also integrated. This is done so that data from previous years can easily be included in the research.

Year	**Number of stations**
2012	100
2013	110
2014	163
2015	159
2016	252
2017	252
Number of stations used for analysis	**261**

Table 5: Number of stations per year

Throughout the years the locations and names of the different stations have slightly changed. For the locations the latest known coordinates of the stations have been used. The spelling and writing of the station names has been standardized.

4.2.2 Preparing the trip data

Just like the stations the format and layout of the individual trips has changed over time. It was determined to make one standardized format to which all the data was transformed. This standardized format helps to easily combine and compare data from different years or months. For each individual trip nine different characteristics have been specified that can clearly identify each individual trip. The characteristics that specify every trip are:

- ID - Each trip has a unique ID
- Code Station (Begin) - The code of the station where the trip started
- Name Station (Begin) - The name of the station where the trip started
- Date (Begin) - The date at which the trip started
- Time (Begin) - The time at which the trip started
- Code Station (End) - The code of the station where the trip ended
- Name Station (End) - The name of the station where the trip ended
- Date (End) - The date at which the trip ended
- Time (End) - The time at which the trip ended

A snapshot of the data and the attributes can be seen in figure 4.

Id	Code Station (Begin)	Name Station (Begin)	Date (Begin)	Time (Begin)	Code Station (End)	Name Station (End)	Date (End)	Time (End)
2	129	Colégio Santa Cruz	01.01.2017	06:48:29	129	Colégio Santa Cruz	01.01.2017	07:47:54
3	276	Fico	01.01.2017	06:49:08	276	Fico	01.01.2017	07:22:42
4	129	Colégio Santa Cruz	01.01.2017	06:54:11	129	Colégio Santa Cruz	01.01.2017	07:48:28
5	19	Parque Ibirapuera - Portão 9	01.01.2017	06:58:53	36	Pedroso Alvarenga	01.01.2017	07:41:19
6	19	Parque Ibirapuera - Portão 9	01.01.2017	06:59:15	36	Pedroso Alvarenga	01.01.2017	07:42:50
7	32	Henrique Martins	01.01.2017	07:14:44	19	Parque Ibirapuera - Portão 9	01.01.2017	07:29:39
8	48	Avenida Ibirapuera	01.01.2017	07:26:32	79	Secretaria da Educação	01.01.2017	11:10:53
9	67	Avenida Faria Lima	01.01.2017	07:37:48	106	Metrô Faria Lima	01.01.2017	16:39:51
11	164	Metrô Faria Lima 2	01.01.2017	07:43:10	36	Pedroso Alvarenga	01.01.2017	07:50:40
12	27	Alameda Casa Branca	01.01.2017	07:45:05	98	James Watt	01.01.2017	08:20:15

Figure 4: Snapshot of data

4.2.3 Determining travel times between OD-pairs

To determine the relation between travel time and number of trips the travel time by bike between different OD-pairs has to be calculated. Using the road/bike network from CEM the distance and travel times between the different OD-pairs has to be determined. Using the shortest distance between the origin and destination is not a good metric for our analysis because different factors that influence the bicycle usage and travel time are neglected.

The travel times are influenced by various factors such as gradients, quality of the infrastructure, availability of the infrastructure (e.g. separated bike lanes), number of turns etc . Using the approach from [Broach et al., 2012], [Pritchard et al., 2019] developed a model that takes into account these various factors for determining the travel time by bike in São Paulo. This model is very suitable for this research because it has been especially calibrated to the characteristics of São Paulo. With this model it was then possible to determine the travel time between the 261 stations. The travel time has been calculated for the shortest path between each origin and destination.

It is important to mention that during the analysis phase two time periods will be used. The first is the rental period. This time represents the interval between the moment the customer picked the bicycle up and returned it. The other time is the travel time, as described the travel time is the theoretical shortest time between two stations.

4.2.4 Calculating the land use mix

Studies have found that the land use around a bicycle sharing station can influence the usage of the system (e.g. [Mateo-Babiano et al., 2016] and [Seskin et al., 1996]). Seskin et al. states that a high land use mixture around stations can positively influence the bicycle usage. Therefore it is necessary to determine the land use mix around each station.

The Secretaria Municipal de Finanças differentiates between 15 different land use types. For the calculation of the land use mixture these elements are grouped into six categories. These six categories are described below and can be seen in figure 5.

1. Education
2. Residential
   - Residential horizontal low standard
   - Residential horizontal medium/high standard
   - Residential vertical low standard
   - Residential vertical medium/high standard
   - Residential
3. Commercial
   - Residential and commercial
   - Commercial and industry
   - Industry and warehouses

Commercial and services

4. Public
   - Public space
   - Vacant land
5. Others
   - No dominant use
   - Garage
   - Others
6. Unknown

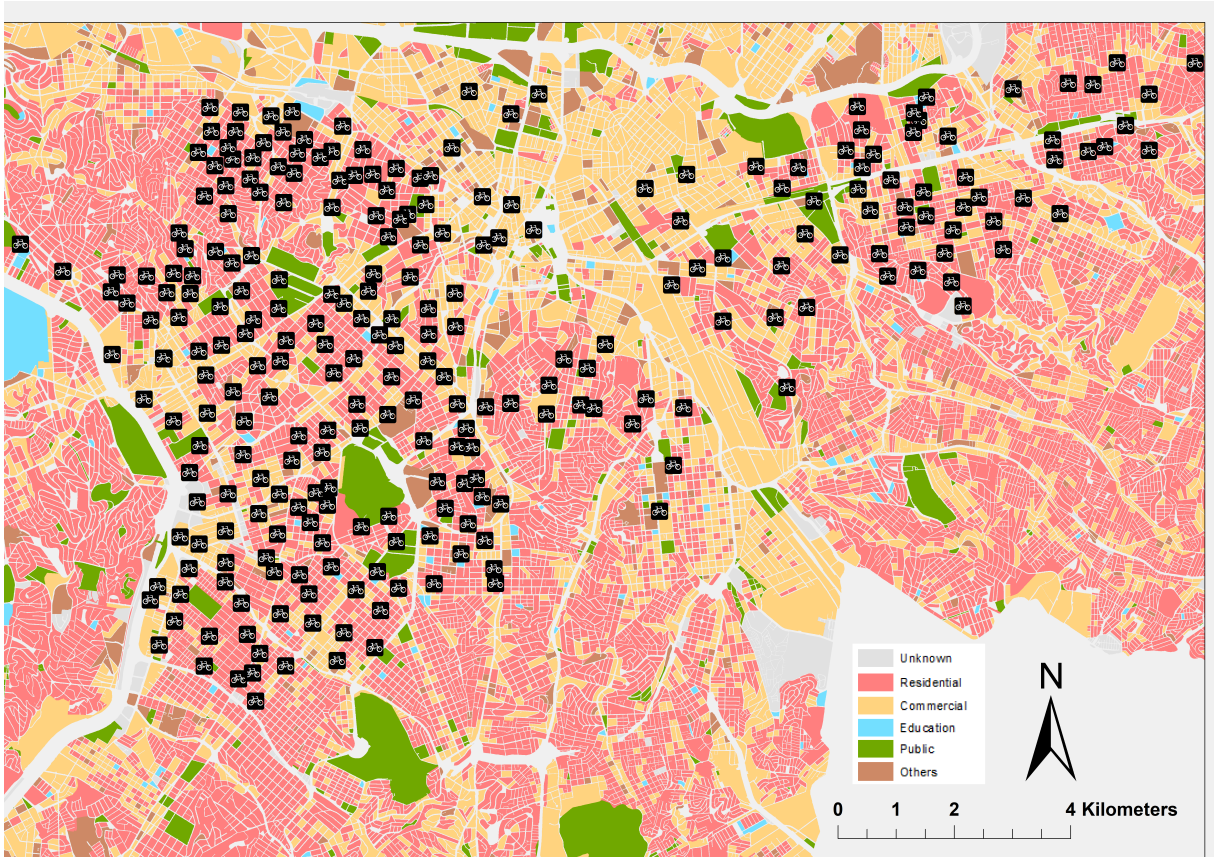


Figure 5: Land use around the system [Secretaria Municipal de Financas 2016]

As measure for the diversity in land-use the 'Land use mixture index' will be used. The land use mixture index determines the variety of different land use categories in a defined area. The defined areas that will be used for this research are the service areas around the stations. The service area is the area that can be be reached within a predefined time interval, starting at the PBSS-station. For the calculation of the land use mix four different service areas are determined. The first two are the areas that can be reached within five and ten minutes cycling from each station. The other two are the areas from which the station can be reached within five respectively ten minutes. The service areas can be calculated using the same model that was used to determine the travel times between origin and destination.

The land use mixture index determines the variety of different land use categories in a defined area. Other studies, [Frank et al., 2006], have also used formula 4.1 to determine the land use mix of an area. The formula for the calculation of the land use mixture is:

$$Land\ use\ Mix_i = \frac{-\sum P_i * ln(P_i)}{ln(k)} \tag{4.1}$$

Wherein P_i is the share of land use category i in one service area and k is the total number of land use categories (six in our case). The outcomes of this equation can vary between 0 and 1, where 0 means that only one land use type is present and 1 means that all six categories are evenly represented in an area. Using the information of the service areas per station the land use mix around each station is determined.

4.2.5 Population density per service area

In the analysis the number of trips is the most relevant variable. Because the number of trips that can depart from a station might be influenced by the number of people living in the surrounding area the population density within the service areas of the origin and destination have been calculated. The population density for each station has been calculated using formula 4.2. The grey area in figure 6 illustrates the service area around a station. For the population density the "5-minutes towards" service area is used. This area has been chosen because it is assumed that five minutes is an acceptable walking distance to the bike stations.

$$Population\ density_{service\ area} = \sum Population\ density_{Census\ area_i} * (\frac{F_i}{Census\ area_i})) \tag{4.2}$$

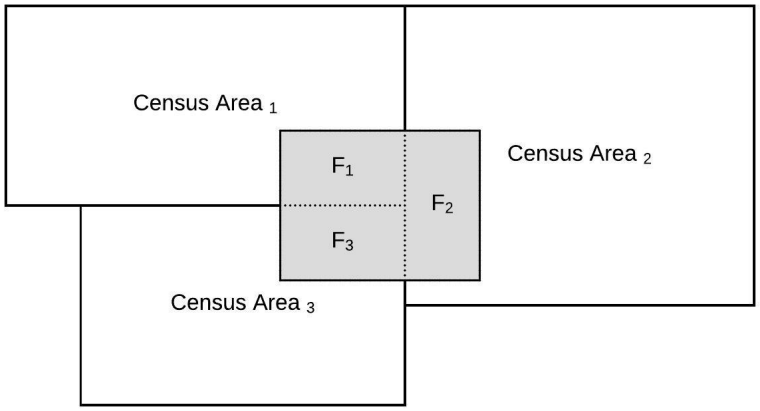


Figure 6: Schema for population density calculation

There are in total six different population density groups. The population density intervals per group have been determined based on the mean (23.427 inhabitants per hectare) and the standard deviation (21.267 inhabitants per hectare) of the population density. The six groups and the distribution of the population density across the service areas of the stations can be seen in figure 7. The blank areas within the figure are not within the service area of a station. For the analysis of the spatial distribution the OD-pairs will be clustered into groups based on the average of the population density around the origin and destination.

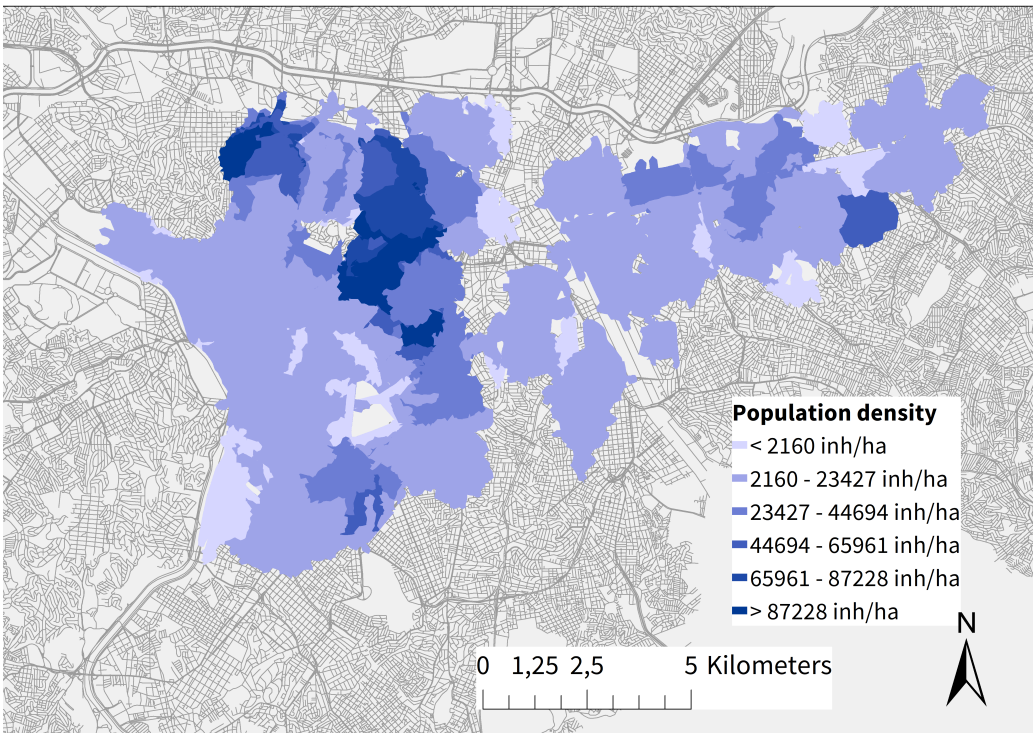


Figure 7: Population density per area [Census (2010)]

4.3 Quality of the data

After the introduction to all the necessary data types and sources it is important to control the quality of the data. This step is necessary to determine if the data is reliable and can be used for the research. The data that is available from the different municipality organizations needs to be considered as complete for this research because there are not enough resources available to check if this data is complete. Therefore this chapter will focus on the quality of the bicycle flow data only.

There are six different data quality dimensions [Askham, 2013]. From this six dimensions four apply to this research:

1. Quantity: How many data do I have?
2. Quality: How good is the data?
3. Uniqueness: How many unique events do I have?
4. Resolution: How detailed is the data?

4.3.1 Quantity and quality of the trip data

As previous mentioned there is data from six different years available for this research. In total there are 520.493 records throughout the years. It has been determined to use only the data from 2017 for various reasons. The first reason is because in the dataset of 2017 there are ten months of continuous data available. In previous years only selected months or days are available. The dataset of 2017 consists of a total number of 234.988 records. The number of records per month, day and time will be presented in the resolution section of this paragraph.

The quality of the data describes the amount of 'bad data', outliers, missing data etc. To assess the quality of the dataset, one therefore has to determine the outliers. For this research it was determined to detect the outliers based on the rental time. This variable has been chosen because it directly reflects the usage of the system. A frequent used method to detect outliers is to calculate the z-score of each rental

time. The z-score describes the number of standard deviations a data point is away from the mean. To calculate the z-score the following formula is used:

$$z_i = \frac{x_i - \overline{x}}{s} \tag{4.3}$$

With x_i the individual rental time, $\overline{x}$ the mean of all rental times and s the standard deviation of the rental times.

After calculating the z-score for each individual trip the trips with a z-score of three or higher have been removed from the dataset. In rental time this means all trips that are longer than 2325 minutes (38,75 hours) are eliminated. In total 4065 data points have been removed from the 2017 dataset. This is share of 1,7% of the total amount of original data.

Given the nature of the data it is not possible to determine if data is missing. It has to be assumed that all rental transactions have been recorded.

4.3.2 Uniqueness and resolution of the data

After eliminating the outliers 230.923 records are left in the dataset. Due to the fact that the beginning of the rental is stored to the second level it can be said that all these 230.923 records are unique records and that there are no duplicates in the data.

The resolution of the data describes how detailed the data is. However how detailed the data is depends on the aggregation level. In table 6 the data is aggregated to months and days. From this table it can be seen that the number of trips in May till July are the lowest and that on Sundays the system was most used in 2017.

Month	**Number of trips**
January	35245
February	37649
March	42304
April	25935
May	12592
June	9337
July	13901
August	18441
September	22745
October	12793

Day	**Number of trips**
Monday	31970
Tuesday	34980
Wednesday	33393
Thursday	33706
Friday	29908
Saturday	28578
Sunday	38489

Table 6: Number of trips per month/day

In the given data the time is stored down to the second level, this level is not suitable for analysis. Therefore the start and end-time of each rental has been rounded to the nearest 30 minutes. The 30-minutes interval has been chosen because it is a good trade-off between efficiency and accuracy. The model that will be described in one of the following chapters incorporates the duration. The duration therefore has also to be rounded. For the rental time it has been determined to round to the nearest five minutes.

4.4 Summary data used for this research

As mentioned the data has been stored in different ways and formats. Synchronizing the data was one of the most important steps before the analysis could start. The preparation has to be done with great care because in a later stage errors can cause problems. However by cleaning and controlling the available data the chances of errors in later stages has been minimized. The five most important things from this chapter are:

- The number of stations involved in this research is 261
- The total number of trips that will be used for analysis is 230.923
- The travel time between the origin and destination of each trip has been determined
- The land use mix of the service areas around the stations has been calculated
- The population density around each origin has been calculated

5 Analysis

This analysis chapter is split into three parts according to the research objectives. First of all the spatial and temporal patterns of the rentals will be investigated. This is done because the literature showed that the spatial and temporal patterns can influence the attraction and generation of the number of trips. For this reason general descriptive statistics of the spatial and temporal patterns will be given. This section is then followed by an analysis of the factors that attract and/or generate trips at the station level for the case of São Paulo. The third section then investigates if these factors are also suitable to forecast the spatial and temporal variations in the number of trips between OD-pairs.

This chapter should help to identify the factors that influence the usage of the PBSS in São Paulo and identify possible spatial and temporal patterns in the usage. It is important to point out and clearly distinguish between trip attraction/generation at the station level and the spatial/temporal variations between OD-pairs in this chapter.

5.1 Spatial and temporal patterns

In this section the focus will be on the number of trips, the duration of the trips and when the trips started. These three elements are important to identify spatial and temporal patterns in the number of trips between OD-pairs.

5.1.1 Trip duration and moments when trips start/end

As previous mentioned the data can be aggregated on different levels. Figure 8 shows the number of trips per day during the week, during the weekend and the total number of trips. From this figure clear patterns can be observed. During the week there are two peaks where most bikes have been picked up. During the weekend the number of trips slowly increases, then stays the same and then decreases again. During the week most of the trips started between 18:05 and 18:25 and ended between 18:25 and 18:50.

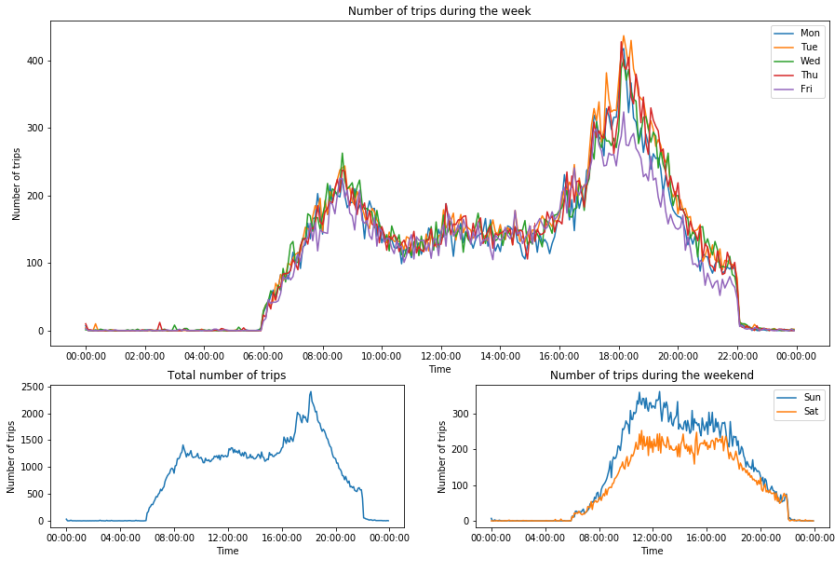
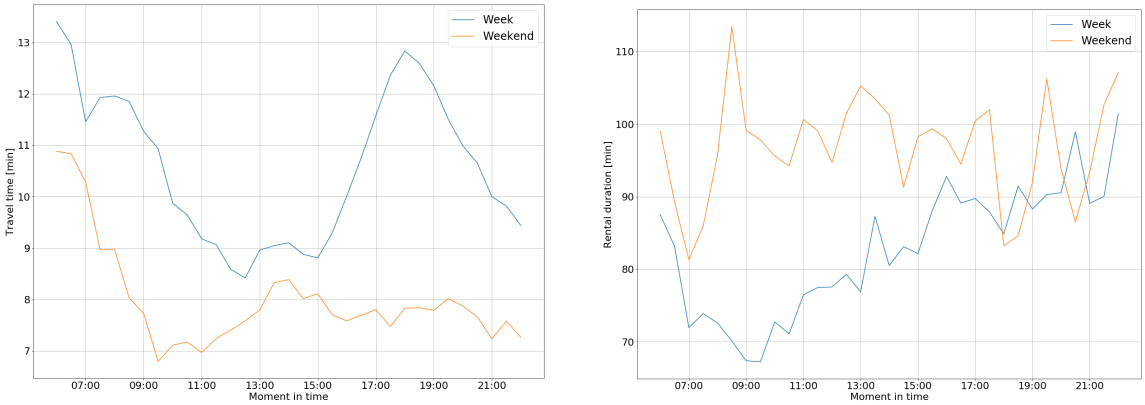


Figure 8: Number of trips

Important: For the following analyses the number of trips that have been made during the week and during the weekend are separately analysed because a clear difference can be seen. In 2017 during the week a total of 145.918 trips took place and during the weekend 59.902.

During the week the mean rental period is 82 minutes. During the weekend the average rental period is slightly higher with 97 minutes. In both cases there occurred one rental transaction that took longer than 38 hours. During the week more than 50% of all trips is shorter than 20 minutes. During the week the median rental time is 10 minutes and during the weekend the median rental time is also 10 minutes.

Figure 9 shows the average travel time and rental duration per time stamp for week days and weekends. As figure 9a shows the travel time during the weekend is always lower than during the week. This means during the weekend trips between OD-pairs that are close are more common. The same figure also shows that the longest trips, distance wise, are made in the early mornings of the week and during the evening rush hour. During the weekend people use the bikes for a longer period of time. Analysing the rental time during the week it can be seen that in the morning people use the bike the shortest. Also a continuous increase in rental duration from the morning until the evening can be obtained during the week.



(a) Travel time per time stamp

(b) Rental time per timestamp

Figure 9: Travel and rental time per time stamp

5.1.2 Number of trips per station

Another level on which the data can be aggregated is the station level. The following figure and tables show the stations where the most trips start and end.

From figure 10 and table 7 it can be seen that in both cases, week and weekend, the stations that are located close to 'Parque Ibirapuera' are frequently used. Especially during weekends these stations have a high demand, probably people use the bikes to cycle around in the park.

The metro station 'Faria Lima' is a popular start- and end-point of rentals because in its proximity there are many areas with commercial land use. Figure 11 zooms in to this specific area. A special thing about the metro station 'Faria Lima' is that there are two stations where bicycles can be rented directly next to the metro station. From a field trip to 'Faria Lima' two interesting things have been obtained. First in the surrounding areas there is a lot of bicycle infrastructure available and the area is relatively flat. Secondly there is a depot of bicycles, so the demand for bikes can always be met.

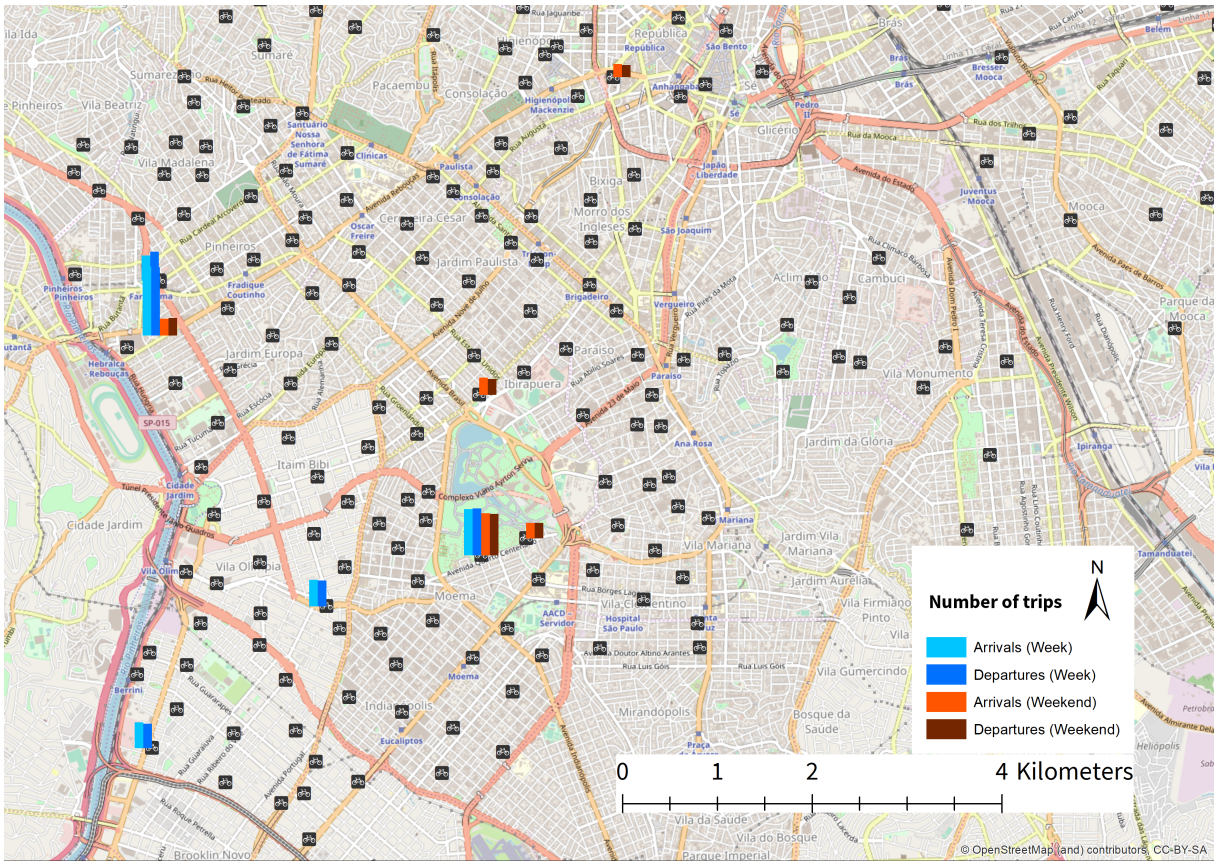


Figure 10: Most frequent used stations

Week		**Weekend**	
Station	Arrivals	Station	Arrivals
Metrô Faria Lima	7985	Parque do Ibirapuera Portão 06	4874
Parque do Ibirapuera Portão 06	5371	Parque do Ibirapuera Portão 09	2013
Faculdade Insper	3059	Metrô Faria Lima	1982
James Watt	2983	Parque do Ibirapuera Portão 05	1806
Metrô Faria Lima 2	2943	Praca Roosevelt	1492
Station	Departures	Station	Departures
Metrô Faria Lima	7928	Parque do Ibirapuera Portão 06	4796
Parque do Ibirapuera Portão 06	5463	Metrô Faria Lima	2035
Metrô Faria Lima 2	3374	Parque do Ibirapuera Portão 09	1849
Faculdade Insper	2998	Parque do Ibirapuera Portão 05	1798
James Watt	2824	Praca Roosevelt	1391

Table 7: Stations with the most arriving and departing trips

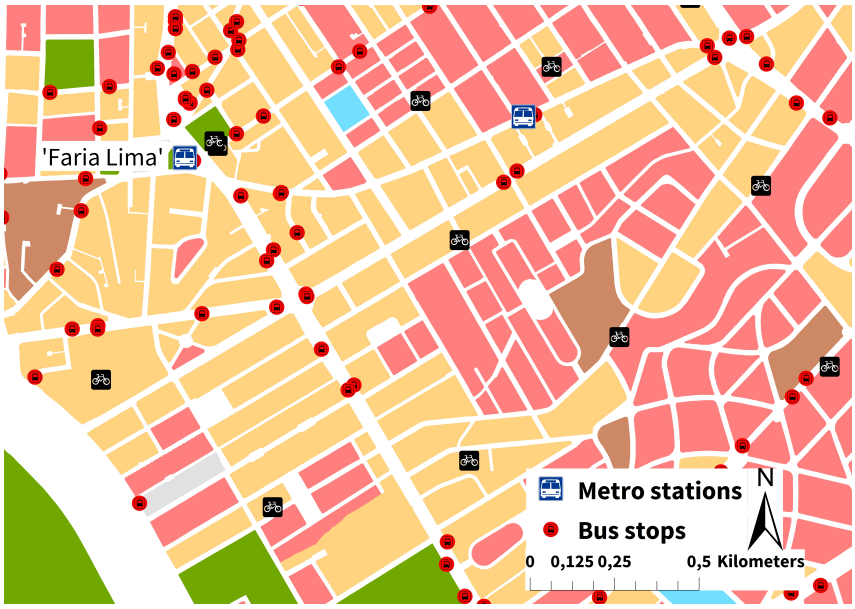


Figure 11: Land use around metro station 'Faria Lima'

5.1.3 Number of trips per OD-pair

With 261 stations that are included in the analysis there are 68.121 possible OD-pairs. However the data from 2017 includes 'only' data for 21.566 OD-pairs. There are 249 OD-pairs with the same origin and destination. From the twenty OD-pairs with the most trips only two have *not* the same origin and destination. The OD-pair with the most trips is the '46-46'-pair. Station number 46 is 'Parque do Ibirapuera- Portão 06', as it can be seen in figure 10 a station with a lot of trips during week days and weekends.

For forecasting temporal variations in the usage of the system the number of records per OD-pair is important. It stands out that from the 21.566 unique pairs, 6.388 OD-pairs only have one record. Integrating these OD-pairs into the model is not suitable because it is impossible to forecast this one trip.

Important: It has been determined that from this step on only OD-pairs with at least five trips or more are used. This means that only data from 8.327 OD-pairs, the number of pairs with five or more trips, will be used for the remaining of this research.

For the research it is interesting to see the number of trips between OD-pairs in relation to the time. Therefore the number of trips between OD-pairs have been plotted for morning and evening peaks during the week. The results can be seen in figures 12 and 13. Again it can be seen that the OD-pairs around the metro station 'Faria Lima' have the highest number of trips. Also interesting to see is that during the evening peak the number of trips is higher. This is in line with the results from figure 8. From further analysis it has been discovered that in the morning most trips depart from 'Faria Lima' and during the evening arrive there.

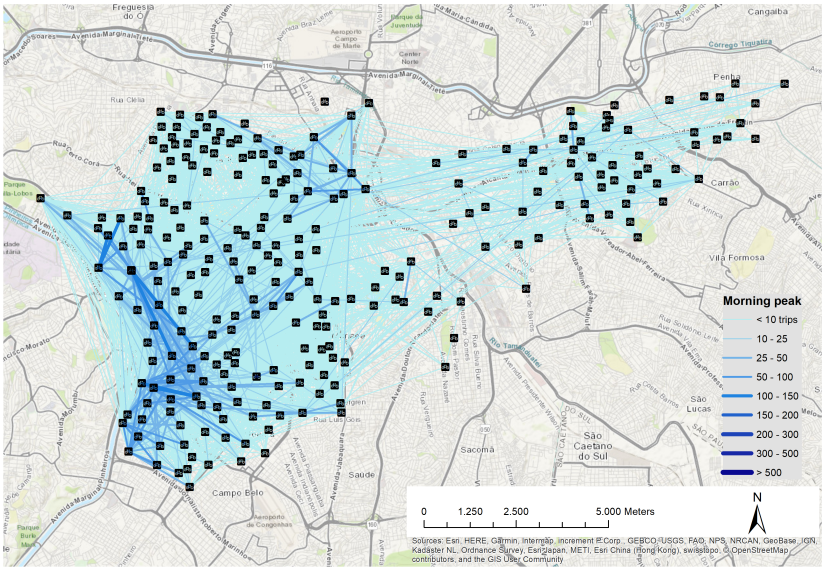


Figure 12: Trips per OD-pair during the morning peak (Week)

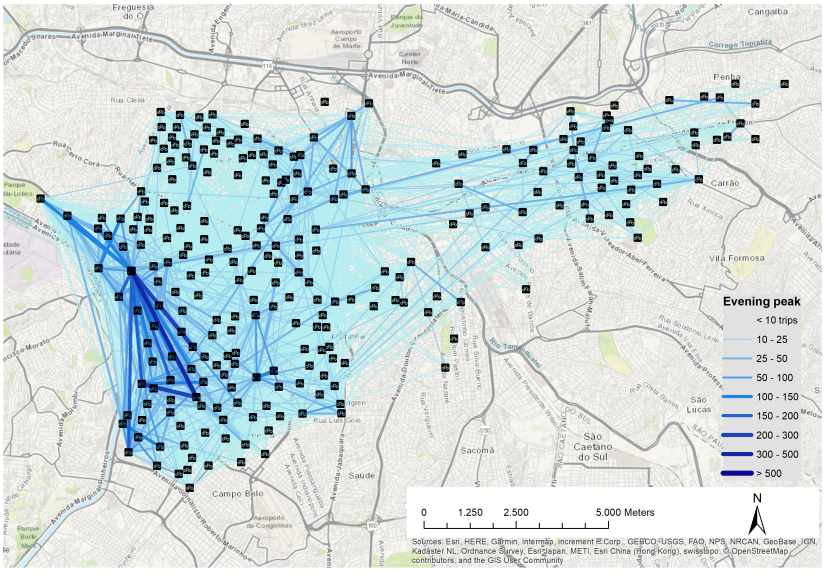


Figure 13: Trips per OD-pair during the evening peak (Week)

5.2 Factors that attract or generate trips per station

As pointed out in chapter 2 the factors that influence the trip generation and attraction are:

- Infrastructure (Availability and quality)
- Land use (Type and diversity)
- City characteristics (Size, topography and climate)
- Connection to other transport modes (Quality of mode and distance to mode)
- Access to the system (Distance and registration)

Except the influence of the city characteristics all the factors will be analysed in this section. The city characteristics will be analysed in the next section that focusses on forecasting the spatial and temporal variations.

To identify possible relations between the characteristics and the number of departing and arriving trips trips per station linear regression models are set up. These models have two purposes for the research, first of all they allow one to determine how good the predictor can predict the dependent variable, in our case the number of arriving and departing trips. The other purpose is more interesting for forecasting the spatial and temporal variations. The model outputs can potentially be used to predict the number of trips between OD-pairs. The linear regression model tries to fit a straight line through the data-points. The R-square value, an output of the linear regression model, indicates how good the fitted line matches the data-points. A higher R-square value, close to 1,00, is desirable. The other important output is the p-value, this value indicates if the predictor is significant or not. For this research, predictors with a p-value lower than or equal to 0,05 are considered significant.

5.2.1 Influence of the surrounding *infrastructure*

The quality and availability of bicycle infrastructure has been pointed out by researchers like e.g. [Cleland and Walton, 2004] to have an important impact on why people cycle. For this reason it was checked if this also applies for the case of São Paulo. There are three 'types' of bicycle infrastructure in São Paulo, they can be seen in figure 14. Table 8 shows the total amount of bicycle infrastructure in the municipality of São Paulo and the amount in proximity to the system. The most common are the 'Ciclofaixa' and 'Ciclovia'. The difference between these two types is that the 'Ciclofaixa' is marked with red paint and the 'Ciclovia' by rubber markers that are attached to the asphalt. For the analysis all three types are combined. A map with all the bicycle infrastructure in São Paulo can be seen in figure 40 in the appendix.

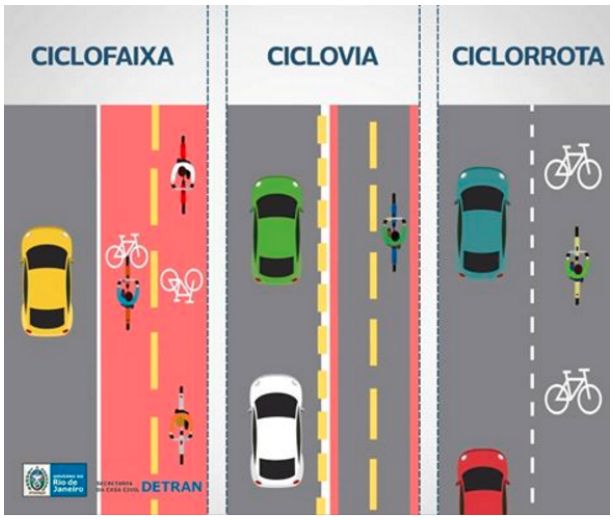


Figure 14: Types of bicycle paths

Infrastructure type	Amount in municipality of São Paulo	Amount in proximity to the system
Ciclofaixa	375 km	144 km
Ciclovia	139 km	64 km
Ciclorota	92 m	-

Table 8: Amount of available bicycle infrastructure

As discovered in section 5.1.1 the most trips take 10 minutes. Therefore the amount of infrastructure within the "10 minutes" service area has been calculated. For the regression models the dependent variables are the number of arriving/ departing trips per station during the week and the number of arriving/ departing trips per station during the weekend. The independent variable is the total amount of bicycle infrastructure , in meters, within the service area.

Figure 15 and table 9 show the results of the regression models. It can be seen from the table that all models are significant. From the figure it can be seen that in all cases if the amount of infrastructure around a station increases the station attracts and generates more trips. This is in line with the expectations from the literature review.

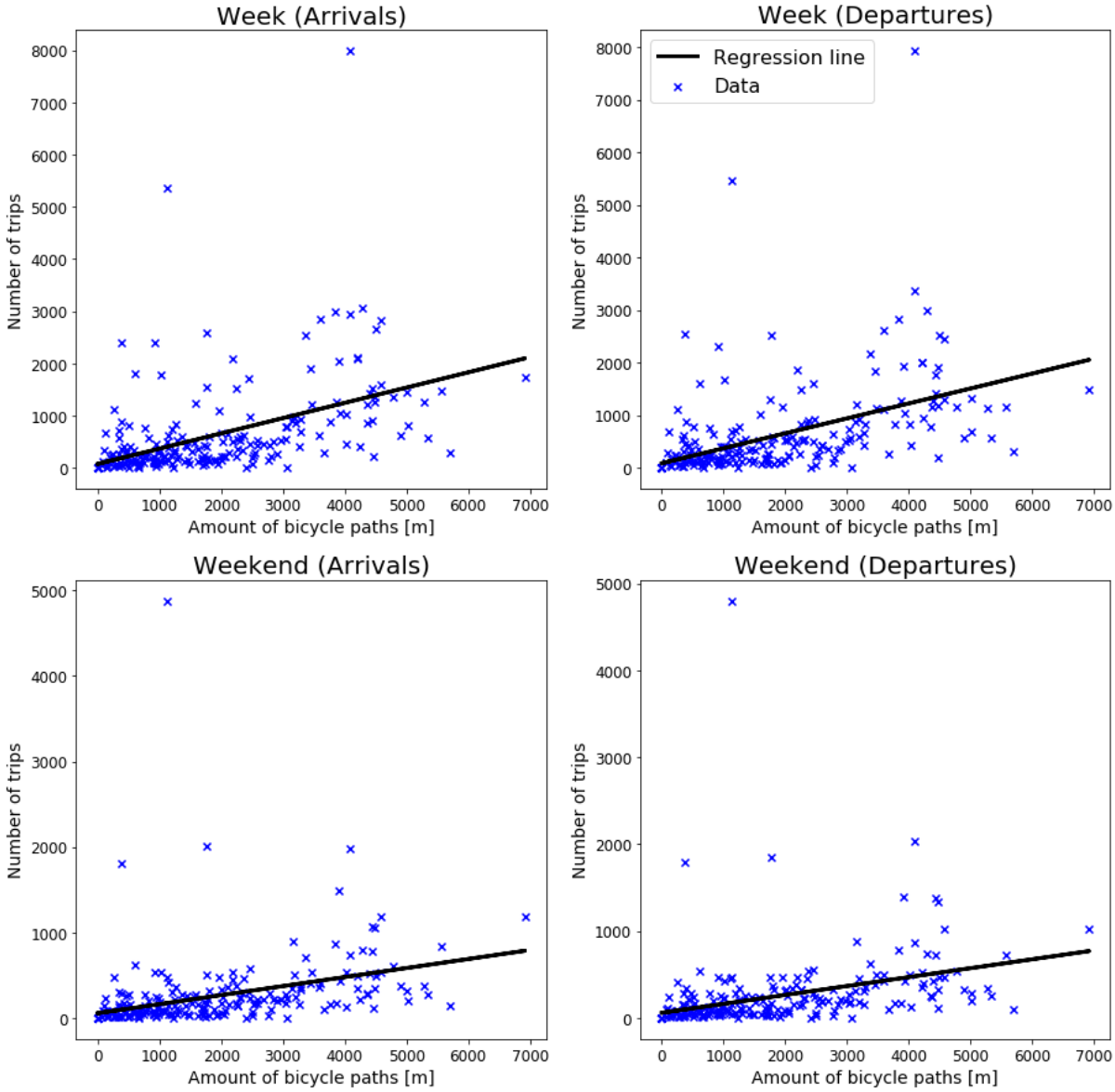


Figure 15: Relation between infrastructure and arrivals/departures

Dependent variable	R-square	p-value
Week (Arrivals)	0,188	0,000
Week (Departures)	0,180	0,000
Weekend (Arrivals)	0,083	0,000
Weekend (Departures)	0,079	0,000

Table 9: Regression results infrastructure and trip attraction/generation per station

5.2.2 Influence of the surrounding *land use*

As identified in the literature review in section 2.1 the land use around stations can influence the number of arriving and departing trips. A mores diverse land use around the stations can cause more arriving or departing trips. To identify if this is the case the land use mix in different service areas has been determined as described in section 4.2.4. The service areas that are used for this research are:

- 5 minutes away from the station
- 10 minutes away from the station
- 5 minutes towards the station
- 10 minutes towards the station

The land use mix in the service areas has been used as independent variables and the dependent variables are the arrivals or departures during the week and weekend. The R-square and p-values of the regression models can be seen in table 10. The results show that all the models are not significant. This means for the case of São Paulo the surrounding land use is not a suitable variable to describe the number of arriving and departing trips of a station.

	Dependent variable	R-square	p-value
Land use mix **5 minutes away** service area	Week (Arrivals)	0,003	0,391
	Week (Departures)	0,006	0,211
	Weekend (Arrivals)	0,004	0,350
	Weekend (Departures)	0,005	0,262
Land use mix **10 minutes away** service area	Week (Arrivals)	0,001	0,598
	Week (Departures)	0,000	0,906
	Weekend (Arrivals)	0,008	0,152
	Weekend (Departures)	0,012	0,085
Land use mix **5 minutes toward** service area	Week (Arrivals)	0,005	0,247
	Week (Departures)	0,005	0,264
	Weekend (Arrivals)	0,003	0,405
	Weekend (Departures)	0,002	0,467
Land use mix **10 minutes toward** service area	Week (Arrivals)	0,012	0,086
	Week (Departures)	0,008	0,153
	Weekend (Arrivals)	0,000	0,812
	Weekend (Departures)	0,000	0,842

Table 10: Regression results land use mix and trip attraction/generation per station

5.2.3 Influence of distance to *transport modes*

The proximity of bike sharing stations to public transport modes can increase the usage according to researches by [Raux et al., 2017], [Zhao and Li, 2017] and [Fuller et al., 2011]. In the case of São Paulo the distance to the nearest bus station, bus terminal, Metro station and CPTM station have been calculated. For the distance calculation the shortest path along the network has been determined in meters. The distance to each of the public transport station has then been used as independent variable in the regression models. The dependent variables are again the number of arriving and departing trips during the week and weekend per station. The results of the regression models can be seen in table 11.

It can be seen that only four models are significant and all include the number of trips during the weekend. The R-square values of the significant models are very low, this means the models do not have a high explanatory power. Interesting to see are the coefficients of the models. They show that if the distance to public transport modes increases the number of trips also increases. This is the opposite of what one would expect from the literature review. The graphs of the significant models can be found in figure 32 in appendix A.1.

	Dependent variable	**R-square**	**p-value**
Distance to **Bus station**	Week (Arrivals)	0,010	0,119
	Week (Departures)	0,009	0,126
	Weekend (Arrivals)	**0,021**	**0,024**
	Weekend (Departures)	**0,021**	**0,024**
Distance to **Bus terminal**	Week (Arrivals)	0,000	0,788
	Week (Departures)	0,001	0,635
	Weekend (Arrivals)	0,005	0,257
	Weekend (Departures)	0,006	0,226
Distance to **Metro station**	Week (Arrivals)	0,003	0,392
	Week (Departures)	0,000	0,863
	Weekend (Arrivals)	0,000	0,966
	Weekend (Departures)	0,001	0,585
Distance to **CPTM station**	Week (Arrivals)	0,000	0,785
	Week (Departures)	0,004	0,329
	Weekend (Arrivals)	**0,016**	**0,044**
	Weekend (Departures)	**0,024**	**0,014**

Table 11: Regression results distance to public transport and trip attraction/generation per station

5.2.4 Influence of *access to the system*

The available data that was provided by CEBRAP and 'Tembici' did not include information about where customers live. This information would be useful to identify the access distance to the system. Researches like the one of [Fuller et al., 2011] found that if people live closer to a station they use the system more often. Instead the population and job density around the stations has been used as proxy for the access distance. It is assumed that higher population or job density is related to distance to the system. These two variables have been used as independent variables in the regression models.

The outcomes of the regression models can be seen in 12. It can be seen that the models including the job density are all significant and the one with the population density is not. Another thing that can be seen from the table is that during the week the job density around the stations has a higher explanatory power than during the weekend. The coefficients from the regression models indicate that if the job density increases the number of trips per station also increases. This is in line with the expectation from the literature.

	Dependent variable	R-square	p-value
Population density	Week (Arrivals)	0,004	0,297
	Week (Departures)	0,000	0,728
	Weekend (Arrivals)	0,004	0,312
	Weekend (Departures)	0,008	0,170
Job density **5 minutes away** service area	Week (Arrivals)	0,091	0,000
	Week (Departures)	0,112	0,000
	Weekend (Arrivals)	0,061	0,000
	Weekend (Departures)	0,071	0,000
Job density **10 minutes away** service area	Week (Arrivals)	0,072	0,000
	Week (Departures)	0,097	0,000
	Weekend (Arrivals)	0,068	0,000
	Weekend (Departures)	0,082	0,000
Job density **5 minutes toward** service area	Week (Arrivals)	0,114	0,000
	Week (Departures)	0,119	0,000
	Weekend (Arrivals)	0,051	0,000
	Weekend (Departures)	0,053	0,000
Job density **10 minutes toward** service area	Week (Arrivals)	0,118	0,000
	Week (Departures)	0,120	0,000
	Weekend (Arrivals)	0,073	0,000
	Weekend (Departures)	0,072	0,000

Table 12: Regression results population and job density and trip attraction/generation per station

5.3 Factors influencing spatial variations

In the previous sections the focus was on trip attraction and generation at the station level. This section will take a closer look at which factors influence the spatial distribution of the trips. Therefore the unit of analysis is the number of trips between an OD-pair instead of the number of arriving or departing trips per station. The same factors as in the previous section and as mentioned in the literature review will be used adjusted to OD-pairs. Also some additional factors that can only be measured for OD-pairs will be analysed.

As mentioned in section 4.2.5 the OD-pairs will be grouped based on the population density around the origin and destination. The purpose of this grouping is to reduce the variance of the dependent variable. By reducing the variance and analysing the data by population groups more accurate statements about the data can be made. The regression models have been set up for the number of trips during the week and weekends separately. These models are then separated again by population density group, that results in twelve different regression models per factor that might influence the number of trips per OD-pair.

5.3.1 Relation between *travel time* and number of trips

As [Polat, 2012] pointed out the number of trips is influenced by the travel time/distance. The literature shows that if the travel time/ distance between origin and destination increases the number of trips decreases. The travel time between OD-pairs is related to the infrastructure and the city characteristics. The model that has been developed by [Pritchard et al., 2019] allows one to calculate the travel time along the shortest path between two stations. For the travel time the model takes the available infrastructure and slope along the path.

Linear regression models with the number of trips as dependent variable and the travel time as independent variable have been set up, to test if this relation also applies for the PBSS in São Paulo. An example of the results can be seen in figure 16. This figure shows the data for the number of trips during the week for OD-pairs with an average population density between 23.427 and 44.694 inh/hec around origin and destination.

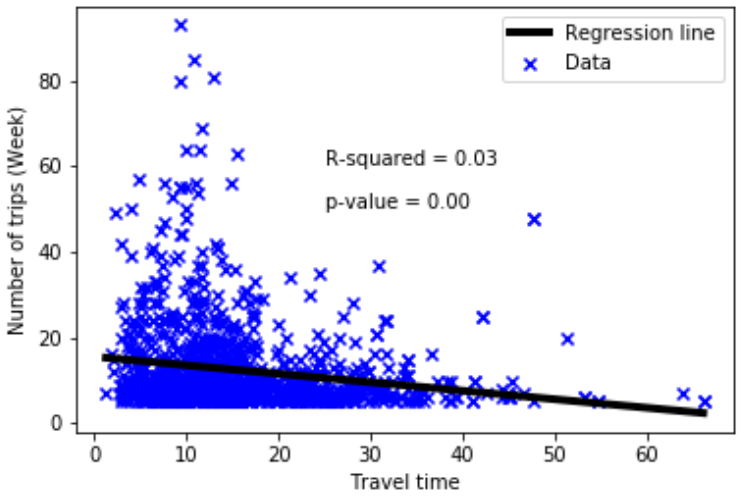


Figure 16: Relation between travel time and number of trips

From figure 16 multiple things can be obtained. First of all the regression line has a negative gradient which means that if the travel time increases the number of trips decreases. The next thing is that the variance in the number of trips between 5 and 20 minutes travel time is very large. The number of trips per OD-pair varies within this time interval between 5 and 93 trips. The p-value of the regression model indicates that the travel time is a significant predictor, however only 3% of the variance in the data is explained by the regression model which is very low and not an acceptable value. The graphs and regression models for the weekends and the other population densities look similar. Their R-square value does not get higher than 10%. For this reason a closer look at the distance decay functions for the different population density groups has been taken.

From figure 17 clear differences between the population density groups can be obtained. The biggest difference can be seen between OD-pairs with a population density higher than 87.228 inh/ha and OD-pairs with an average population density between 44.694 and 65.961 inh/ha. The OD-pairs with the highest average population density have the shortest trips, 50% of their trips are shorter than 10 minutes. The distance decay functions for the weekend for the groups look similar and no big differences can be seen between weekends and week days.

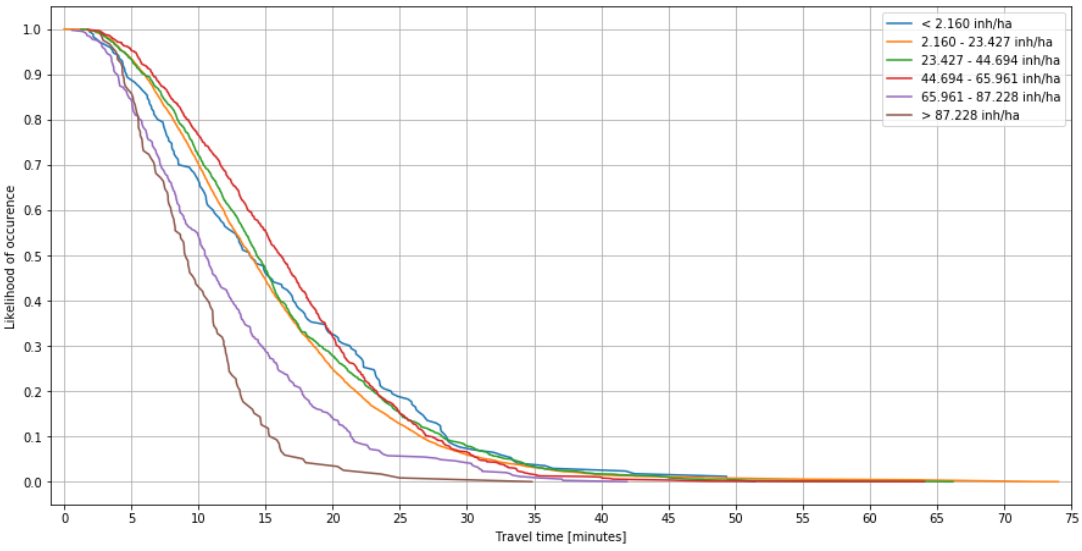


Figure 17: Distance decay function (Week)

To describe these distance decay functions different mathematical functions can be used. The 'most widespread are the exponential and the power function' [Martínez and Viegas, 2013]. The general form of the exponential functions that are fitted to the distance decay functions is presented by equation 5.1. For this function three parameters have to be fitted, the fitted parameters per distance decay function can be seen in appendix A.2. In their article Martinez et al. also present another functions which is according to them more suitable to describe a distance decay function. They propose to use 'Richard's function', this function was initially proposed for the field of botany by Richards in 1959. The function is shown in equation 5.2. In our case C is set to 0 and K to 1. Like the parameters of the exponential function the fitted parameters of the Richard's function can be found in appendix A.2.

$$f(x) = a * e^{b} + c \tag{5.1}$$

$$f(x) = C + \frac{K - C}{\left(1 + Qe^{-B(x-M)}\right)^{\frac{1}{v}}} \tag{5.2}$$

Figure 18 shows an example of the exponential fit and the Richard's function. It can be seen that the Richard's function describes the data more accurate. In all cases the root mean squared error of the Richard's function is smaller, see table 22 in appendix A.2.2.

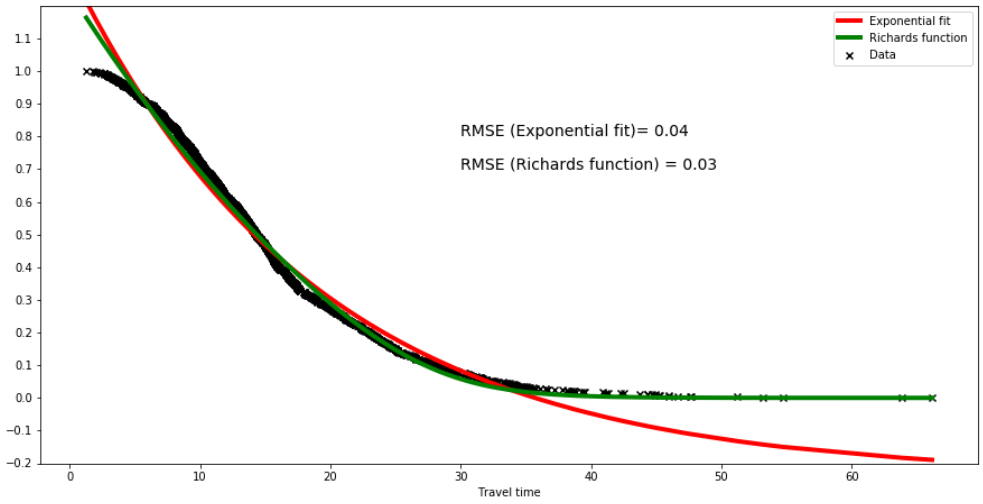


Figure 18: Distance decay function and fits

5.3.2 Relation between *available infrastructure* and number of trips

In section 5.2 only the surrounding infrastructure was taken into account. Because this section looks at OD-pairs it is now possible to determine the share of bicycle infrastructure along the route. To identify the impact of the infrastructure on the number of trips the route along the shortest path between the origin and destination has been determined. The next step was to identify the share of the route that has bicycle infrastructure. The share of bicycle infrastructure per OD-pair has then be used as independent variable. In case of OD-pairs with the same origin and destination the share of bicycle infrastructure within the 15 minutes service areas has been determined.

Again all twelve regression models, one per population density and then separately for week days and weekends, have been run and only four are significant. The models that are significant do not have a R-square value higher than 0,02. Again the data is very scattered as the example in figure 19 shows.

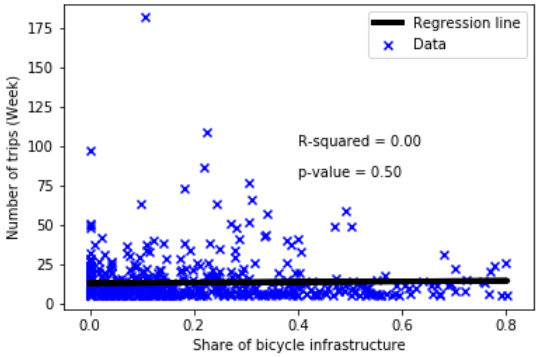


Figure 19: Relation between available infrastructure and the number of trips

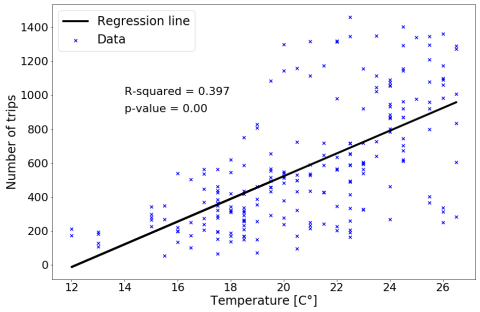
5.3.3 Relation between *weather* and number of trips

As pointed out in the literature review by [McLoughlin et al., 2012] the weather influences the number of trips within a PBSS. For this sub-section a closer look at the total number of trips within the system will be taken, not at the OD-pair or station level. This is done because the weather data is not detailed enough to obtain the temperature of each individual station location.

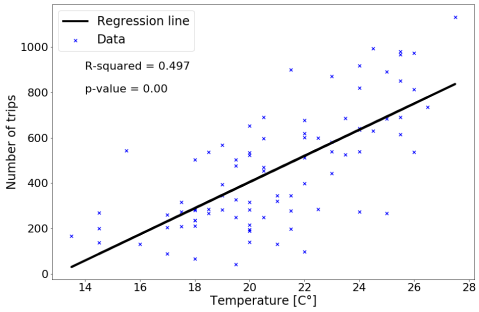
The weather factors that are used as independent variables in the regression model are the average temperature per day and the daily amount of rain. For the analysis four different 'groups' of trips have been created. The four groups are:

- Trips with same origin and destination during the week
- Trips with same origin and destination during the weekend
- Trips with different origin and destination during the week
- Trips with different origin and destination during the weekend

Only the regression models using the temperature to predict the number of trips within the system are significant. The models that use the amount of rain have all a p-value higher than 0,05. Table 13 and figure 20 show the results of the regression models. The regression models that use the temperature have a positive slope which is in line with the expectations from the literature. However again the data points are wide spread. The exact coefficients of the regression models can be found in appendix A.4.



(a) Trips with different origin and destination (Week)



(b) Trips with different origin and destination (Weekend)

Figure 20: Regression models with temperature as independent variable

Independent variable	R-square	p-value
Same origin and destination (Week)	0,279	0,000
Same origin and destination (Weekend)	0,410	0,000
Different origin and destination (Week)	0,397	0,000
Different origin and destination (Weekend)	0,497	0,000

Table 13: Regression results temperature and number of trips

5.3.4 Relation between *change in job accessibility* and number of trips

In previous sections the job density around the stations has been used. For the analysis of the spatial variation the job accessibility by bike and public transport has been used as independent variable. This has been done because the job-accessibility is a more suitable measure for spatial interactions. The model of [Pritchard et al., 2019] allows one to calculate the job accessibility using a combination of bike and public transport. This analysis was only performed for trips during the week because the model for the accessibility calculation does not take into account weekdays.

For the different OD-pairs the change in job accessibility from origin to destination has been determined. An increase means that at the destination of the trip the job accessibility is higher. As in the previous sections it has been tried to fit a linear regression model to the data. The results of the two regression models can be seen in figure 37 in the appendix. It can be seen that the models are significant but have a low explanatory power.

Similar to the relation between travel time and the number of trips a closer look is taken at the exponential functions of the change in job accessibility. From figure 21 differences between the population density areas can be obtained. While for areas with a population density higher than 87.228 inh/ha the likelihood of an increase in job accessibility is about 65%, the likelihood for an increase in job accessibility is only about 40% for areas with an population density between 2.160 and 23.427 inh/ha. Figure 22 shows the change in job accessibility for the 10 minutes service areas around the stations. In this figure the differences between the population density areas become even more clear.

Again, as it has been done for the travel time, an exponential curve has been fitted to the different lines. This time only the 'Richard's function' has been fitted because as section 5.3.1 has shown this function describes the exponential decrease more accurate. The parameters and the residuals can be found in appendix A.3.

The results of this analysis show that in denser populated areas trips to areas with a higher job accessibility are more common.

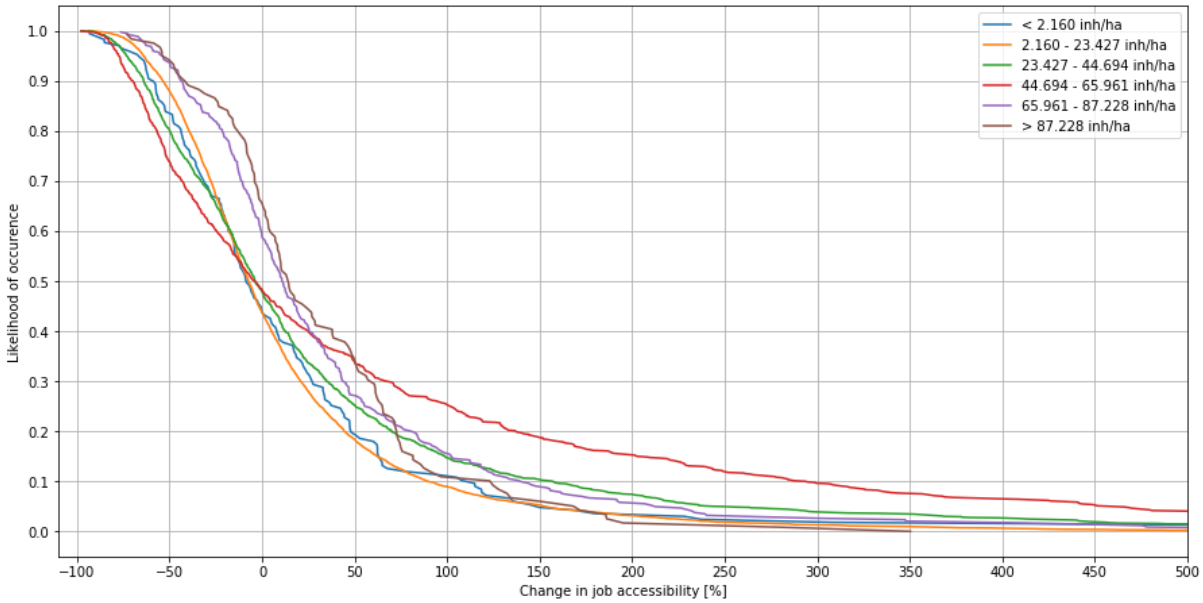


Figure 21: Exponential decay function of the job accessibility (5 minutes service area)

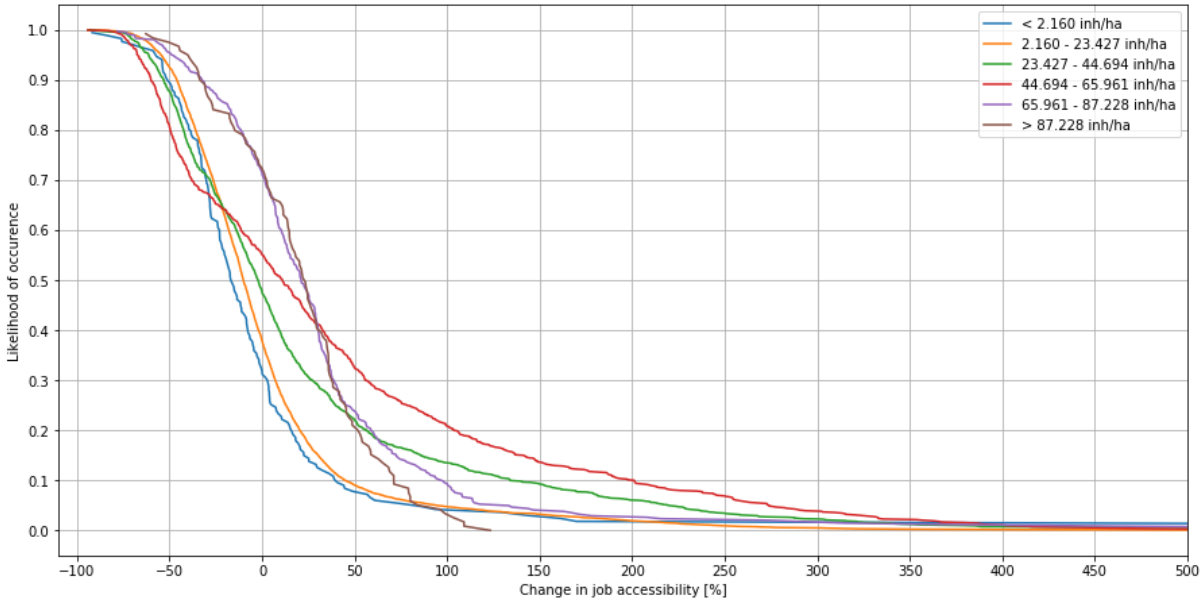


Figure 22: Exponential decay function of the job accessibility (10 minutes service area)

5.3.5 Relation between *other factors* and number of trips

From the factors that are mentioned on page 17 in the literature review until now only a few have been described. This is due to the fact that the models that include the other factors are all not significant or have a very low R-squared value. In this section therefore only a very brief description of the results will be given.

Distance to public transport

To identify the impact of the distance to the transport mode, the average distance at the origin and at the destination has been taken. Two graphs for the four transport modes and trips during the week and weekend can be seen in appendix A.1 figure 33 and 34. These graphs show that there is no regression

model with a R-square value higher than 0%. The models have also been created for the different population density groups but similar results were obtained.

Slope between origin and destination

The model of [Pritchard et al., 2019] also allows one to determine the maximum slope one has to cycle from origin to destination. The slope is one of the city characteristics. As it can be seen in figure 35 in the appendix the models are significant however their R-square values are very low.

Change in land use

To obtain the influence of the land use on the spatial variation the relative change in land use mix from the origin to the destination has been calculated. A positive value therefore means an increase in land use mix, this means at the destination a more diverse land use is present. The land use mix has been calculated for the 5 and 10 minutes service areas around the stations. The OD-pairs with the same origin and destination have been excluded because no change in land use is measurable. The results can be found in figure 36 in the appendix. Only the model for the weekend that uses the change in the 5 minutes service area as independent variable is significant, but it has a R-square value close to 0%.

5.3.6 Relation between *multiple factors* and the spatial variation

In previous sections only single variables have been used to set up simple linear regression models. In this section a factor analysis is performed to identify possible relations within variables. The idea behind the factor analysis is to investigate relationships between different variables that are not easy to identify because the variables are associated with a latent variable [Rahn, 2018]. Before the results are presented it is important to know some variables of the factor analysis. The factors have been extracted based on an Eigenvalue greater than 1. For the rotated component matrix the Varimax-maethod has been used. The factors for days during the week and weekend can be seen in table 14 and 15.

Factor	**Variables**			
1	Accessibility change BnR (5 minutes)	Accessibility change BnR (10 minutes)	Accessibility change PuT (5 minutes)	Accessibility change PuT (10 minutes)
2	Distance to bus station	Distance to all PT modes		
3	Change land use (10 minutes)	Distance to bus terminal	Distance to CPTM	
4	Distance to metro	Maximum slope	Travel time	
5	Share of bicycle paths			
6	Change land use (5 minutes)			

Table 14: Extracted factors (Week)

Factor	**Variables**			
1	Accessibility change BnR (5 minutes)	Accessibility change BnR (10 minutes)	Accessibility change PuT (5 minutes)	Accessibility change PuT (10 minutes)
2	Distance to bus station	Distance to all PT modes		
3	Distance to bus terminal	Distance to CPTM		
4	Distance to metro	Maximum slope		
5	Share of bicycle paths	Travel time		
6	Change land use (5 minutes)	Change land use (10 minutes)		

Table 15: Extracted factors (Weekend)

The exact load factors for the different variables for week days and weekends can be seen in appendix A.5. Using this information about which variables are correlated a 'new' variable was calculated. This new variable is the sum of all variables per factor. It was assumed that all variables have the same weight, due to the limited amount of time of the research it was not possible to investigate the different weights of the variables, what would have been desirable.

After the calculation of the 'Factor-variables' the OD-pairs with the 50 highest and lowest values per factor are selected. The average number of trips for the highest 50 OD-pairs are compared with the lowest 50 to discover possible difference. To identify if the means are significant different a T-test is used. The results from the T-test, that can be seen in appendix A.5.1, show that using the 'Factor-variables' the 50 highest OD-pairs have a significant different number of trips than the 50 lowest OD-pairs. This is true for all factors with the exception of factor 5, for week days and weekends. For factor 5 the 50 highest OD-pairs do not have different number of trips than the lowest 50. This 'Factor-variable' has been used as an independent variables in simple linear regression models. Again no high R-square values were obtained.

5.3.7 Methods used for increasing the explanatory power

As described in section 2.7 different clustering techniques were used to increase the explanatory power of the regression models. The previous sections already pointed out that the data of the different OD-pairs is very scattered and therefore hard to predict. However different approaches have been taken to group or categorize the data to fit the regression lines better to the data. In this section the methods that have been used are explained.

The first approach was to make different regression models based on the number of trips, e.g. a regression model for OD-pairs with a number of trips between 5 and 50, 51 and 100 etc. As it can be seen in figure 38 on page 68 about 50% of the OD-pairs have less than 10 trips. Different intervals have been tested also in combination with the population density categories. However this approach did not lead to an increase in the R-square values.

The second approach was to cluster the stations based on their geographical location. The stations that are represented in figure 2 have been organized into groups. Regression models for trips within the clusters and between the clusters have been created. Again this approach did not increase the R-square values.

Due to the fact that clear temporal patterns have been identified models have been set up for morning- and evening-peaks and the period in between. In these models the dependent variable was the number of trips that occurred in the relevant period. This approach also did not result in an increase of the explanatory power of the regression models.

5.4 Summary analysis

The previous sections described the factors that influence the usage of the PBSS in São Paulo and identified spatial and temporal patterns of the usage. First of all a clear difference in the usage between week days and weekends can be obtained. During the weekends the trips that are made within the system are longer and they start and end often at the same station. From Monday until Friday most of the trips are made in the west of the system. This is due to the fact that a lot of commercial activities are located within this area and the station density is higher.

The second part of this chapter identified if the factors that have been observed from the literature also apply to the case of São Paulo. For the trip attraction and generation not all the factors apply to São Paulo. The regression models that used the factors to describe the number of arriving or departing trips have a low R-square value, if they are significant.

The third part of the analysis investigated if the factors that apply to trip attraction and generation can be used to describe the spatial variations in the number of trips. Therefore a closer look at the number of trips between OD-pairs has been taken. From the single variables that have been tested the average daily temperature is most suitable to estimate the number of trips for the whole system.

Table 16 compares the results from the models that were used for the trip attraction and generation and the models that were used for the spatial variation. It can be seen that the factors that can be used to describe the trip attraction and generation at the station level are not always suitable to describe the spatial variations. Vice versa the factors that are used to describe the spatial variation are not always suitable to describe the trip attraction and generation. Only the job density/ accessibility is significant for both cases.

Different grouping and clustering techniques have been applied to increase the explanatory power of the linear regression models, however these techniques did not increase the R-square values of the models. Nonetheless the factor analysis showed that combining different variables to explain the number of trips can be helpful

Variables		**Trip generation and attraction**		**Spatial variation**	
		Significant	**R-square values**	**Significant**	**R-square values**
Infrastructure		Yes	0,188 - 0,079	Not	-
Land use		Not	-	Only weekend (5 minutes)	0,001
City characteristics		*Not tested*		Yes	0,014 & 0,030
Connection to public transport	**Bus station**	Only weekend	0,021	Yes	0,002 & 0,043
	Bus terminal	Not	-	Not	-
	Metro station	Not	-	Only week	0,000
	CPTM station	Only weekend	0,014 & 0,044	Only week	0,003
Access to the system	**Population density**	Not	-		
	Job density	Yes	0,051 - 0,120	Yes	0,001 & 0,002
Travel time		*Not tested*		Yes	0,030
Weather		Yes	0,297 - 0,497		

Table 16: Comparison of the regression results

6 Forecasting the number of trips between OD-pairs

Next to identifying the factors that influence the numbers of trips and discovering spatio-temporal patterns the main focus of this research is the development of a model that can forecast spatial and temporal variations in OD-pairs by using the gained insights in the 'behaviour' of OD-pairs. This chapter will describe the development of this model. To identify the variations the model will be able to forecast the number of trips per OD-pair per time stamp and give an estimate of when the bikes will arrive at the destination. A visual representation of the model in the form of a flowchart can be seen in figure 23. The sections in this chapter will focus on each part of the model.

6.1 Preparing the data

The first section, the one with the blue background in figure 23, is to select and prepare the necessary data. As input all available trip data from 2017 is used that already has been described and analysed in chapter 4 and 5. As previously mentioned to improve the quality of the model and reduce randomness only OD-pairs that have more then five transactions within in the research-period are selected. In contrast with the previous chapters the data from 2012 until 2016 will also be used to set up the model. Only the transactions between OD-pairs that had more than five trips in 2017 are selected from the datasets between 2012 and 2016. This is done to improve the predictions. The reason the data between 2012 and 2016 is added is because we are interested in the temporal patterns. In the previous chapters we were interested in which factors influence the usage. After this all trips that started between 06:00 and 22:00 are selected because these are the operating hours of the system. This time period can also be clearly identified from figure 8.

Summarizing, after adding the data from 2012 until 2016 and the removal of trips that are not within the operating hours, 438.862 trips are left to set up the model. A common practice is to use 90% of the total data to set up the model and the remaining 10% to validate it. In our case 394.976 trips are used to set up the model ad the remaining 43.886 to validate the model.

6.2 Calculating the necessary data

The next step of the model is to summarize the data of each individual OD-pair in a separate dataframe. A 'DataFrame is a 2-dimensional labelled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table.' [Pandas Documentation, 2018]. Each dataframe includes all the necessary information from one specific OD-pair. All the calculations and steps that are described in the following, can be seen in the red area of figure 23.

The first step is to count how many trips happened per date-time-stamp. A date-time-stamp includes the date and time when a trip started e.g. 05.03.2017 06:12:39. These date-time-stamps are rounded to the nearest half hour. This is done for two reasons, the first is that summarizing the data on a half hour level reduces the computing time and the second is that predicting the number of trips on a seconds level is not practical. After the date-time-stamp has been rounded the sum of each individual stamp is calculated. The next step is to calculate the average number of trips per time-stamp. The time-stamp of a trip is only the rounded departure time without the date. In total there are 32 timestamps between 06:00 and 22:00. After this step the average number of trips per time-stamp per OD-pair is known. With this information the total number of trips per OD-pair per day can be determined. This number is necessary to calculate the cumulative probability per time-stamp. All these steps are executed for each of the 8.327 OD-pairs.

The last process in the red area is to calculate the average rental period per time-stamp. This information can be used in the final prediction. Adding information about the average rental period can be an important information to the operator because with this information the operator knows when to expect the bicycles at the destination. Calculating all the necessary information for the OD-pairs beforehand can reduce the running time of the model later if the information from the dataframes are used like a

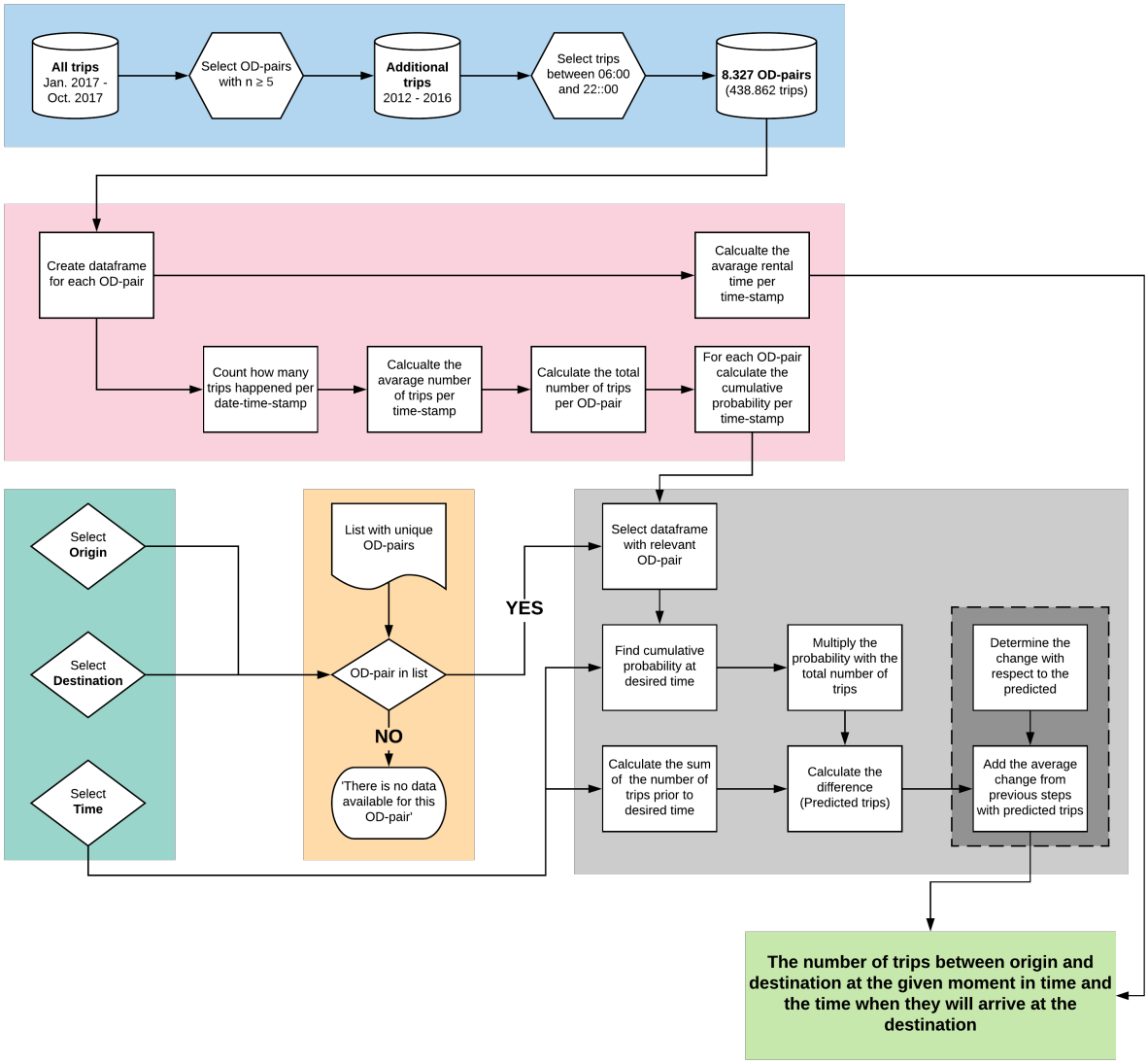


Figure 23: Visualization of the model

database.

6.3 Model input and data selection

As mentioned previously there are 261 stations that are included in this research. This would mean if there are trips between each station there would be 68.121 unique OD-pairs. In 2017 'only' 8.327 of the OD-pairs have had more than five trips.

These individual OD-pairs are summed up into one list, see the yellow area of figure 23. Next to this yellow area is an area with a cyan background. In this step of the model the variables are defined. The three input variables for the model are:

- Code of the origin station
- Code of the destination station
- The moment in time for which to predict the number of trips

Using the origin and destination input one can then determine the OD-pair and check if the desired OD-pair is in the list with unique OD-pairs. If the OD-pair is not included in the list of unique OD-pairs, the model returns the following answer: 'There is no data available for this OD-pair'. However if the OD-pair is included in the list the model selects the relevant dataframe and continues running.

6.4 Forecasting the number of trips

The previous sections focussed on the selection and preparation of the necessary data, this section will describe the steps that are required to calculate the number of trips per time stamp from the historical data and actual usage. For the forecast some things are important, first of all 'In general, regular OD trip desires can be viewed as a repeated process with similar within-day dynamic patterns' [Zhou and Mahmassani, 2007]. This result from Zhou et. al has also been discovered in this research, see figure 8. Therefore two separate models are set up, one for Monday until Friday forecasts and the other that is used for weekends. Another crucial assumption of the model is that the total number of trips for each OD-pair stays constant.

In the development of this research one has thought about incorporating factors around the stations into the model. However as chapter 5 showed the surrounding factors are not suitable to predict the number of trips per station or to forecast spatial variations. It is desirable that the model can react to trends, one of these trends might be changes in temperatures. As it was discovered in section 5.3.3 if the average daily temperature increases the total number of trips within the system also increases. For this reason the general model is run for days with high and low temperatures, more about this in a later section. To react on other trends like events the model includes a part that checks for trends and then adapts the forecast. Zhou et. al used formula 6.1 to define the demand between OD-pairs. The research of Zhou et. al was focussed on predicting real time travel demand. The main difference between this research and the one of Zhou et. al is, that this research identified the deviations caused by different factors. The model of Zhou. et al only recognizes these deviations and then reacts to them.

$$True\ demand = regular\ pattern + structural\ deviations + random\ fluctuations \tag{6.1}$$

All the steps that are required to forecast the number of trips per time stamp per OD-pair can be seen in the grey area of figure 23. The first step is to select the dataframe with the desired OD-pair from the database. The next step is to identify the cumulative probability at the time stamp that has been defined as input. The cumulative probability and the total number of trips are then multiplied to obtain how many trips should have occurred *after* the time stamp. To identify the number of trips during the time stamp one first has to determine the sum of the number of trips prior to the desired time stamp. The difference between the number of trips at the end of the time stamp and the number of trips before the time stamp is the number of trips that will depart within the time stamp and go to the destination. A step by step description of the forecast is given below:

- Step 1: Define origin, destination and time-stamp
- Step 2: Select relevant dataframe from database
- Step 3: Determine cumulative probability at time-stamp
- Step 4: Multiply cumulative probability with the total number of trips
- Step 5: Calculate the sum of trips prior to time-stamp
- Step 6: Subtract the sum of trips prior to the time from the product of cumulative probability and total number of trips
- Step 7: Determine the average difference between observed and predicted values of previous steps.
- Step 8: Add the average difference to the prediction

Until this point the model does not take trends into account, therefore it is a static model based on historical data. The darker grey area surrounded by a dashed line in figure 23 indicates the part where the trends are taken into account. To detect a tendency in the number of trips the predicted number of trips are compared with the actual number of trips. The average difference between observed and predicted of previous steps is then added to the prediction. For example if on average the observed number of trips of the previous two steps was two trips higher than predicted, then these two trips are added to the prediction of the third step. More about calculating the average in section 6.5. Calculating a relative change is not suitable because there are many timestamps with zero trips and it is impossible to determine the relative change with respect to zero. The method of calculating the average difference and then adding it to the prediction can be seen as the 'structural deviations' part from equation 6.1.

The final step is to look up the average rental period for the OD-pair at the time stamp from the database to give an estimate when the bikes will arrive at the destination station.

6.5 Validation number of trips

After setting up the model and the description of it, the next step is to validate the outputs of the model. As previously already mentioned 90% of the complete data is used set up the model and the remaining 10% to validate it. There are two steps of validating, the first is the validation of the 'static' model so without adding the forecast by the trend and the second is the validation of the 'complete' model. This is done to evaluate the accuracy of the regular pattern forecast and to see if the addition of the trend factor increases the performance of the model.

6.5.1 Validation regular pattern

As mentioned two separate models, for week days and weekends, have been set up. The 90% of the data and the remaining 10% are selected randomly. From the 10% portion of the dataset, ten random days during the week and ten during the weekend have been selected. For each day the residuals have been calculated using formula 6.2, with y_i the observed value and $\hat{y}_i$ the predicted value. This step is done ten times to ensure that different parts of the dataset are used for calibration and validation.

Figure 24 and 25 show the density plot of the residuals. The average residual of the week days is is higher than the one of the days during the weekends, but during the weekend the standard deviation of the residuals is higher. Next to the mean and the standard deviation the 95% confidence intervals of the residuals are calculated. From the results in table 17 it can be seen that the confidence interval during the week is a bit smaller, but also that the intervals are both very narrow.

$$Residual = y_i - \hat{y}_i \tag{6.2}$$

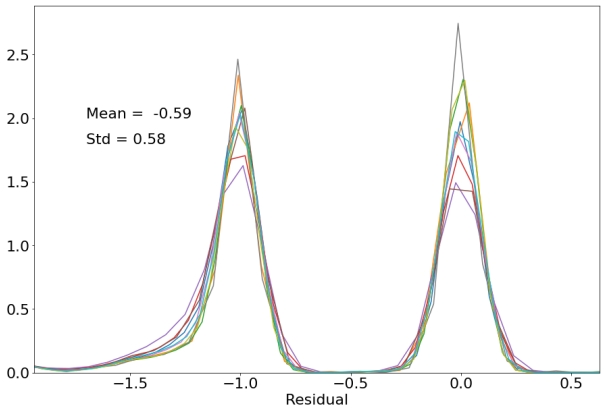


Figure 24: Residuals (Week) - regular pattern only

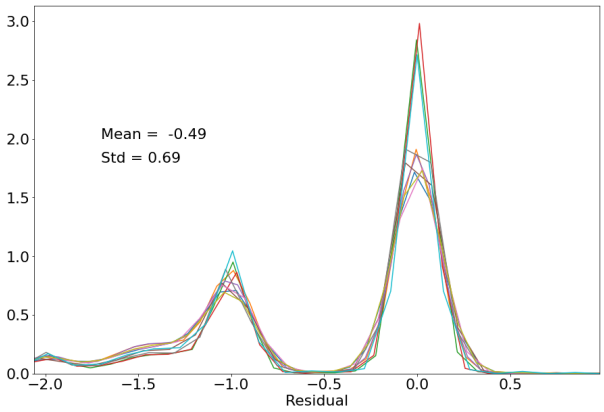


Figure 25: Residuals (Weekend) - regular pattern only

	Upper bound	Lower bound	Width
Weekend	-0,4853	-0,4929	0,0076
Week	-0,5902	-0,5953	0,0051

Table 17: 95% confidence interval of residuals (Only historical data)

6.5.2 Sensitivity analysis of time intervals

The information about how many trips took place between the OD-pair before the desired moment in time will be used to detect trends in the rental pattern and maybe improve the predictions. Until now it is unclear from how many previous time stamps information should be included, for this reason different intervals have been tested.

To identify the most suitable time interval different intervals have been tested and compared based on their average residual. The intervals that have been tested are:

- 30 minutes prior - one time stamp
- 60 minutes prior - two time stamps
- 2 hours prior - four time stamps
- 4 hours prior - eight time stamps
- All time stamps prior to the desired time stamp

The comparison of the average residual for the different intervals show that the information, historical data and observed number of trips, from the previous **one time step** results in the lowest average residuals. Therefore it has been chosen that for the model only information of the previous time stamp is used to improve the model and detect possible trends.

6.5.3 Model validation incorporating information from previous time stamps

As the previous section has shown the information of the previous time stamp will be used to detect trends and possibly improve the forecast. The results of the density plot and confidence interval can be seen in figure 26 and 27 respectively table 18. It can be seen that by using the information from the previous time stamp the average residual is reduced by about 50%.

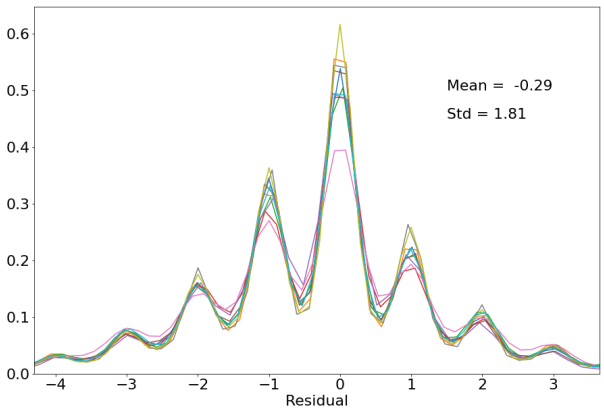


Figure 26: Residuals (Week) - Using information from previous steps

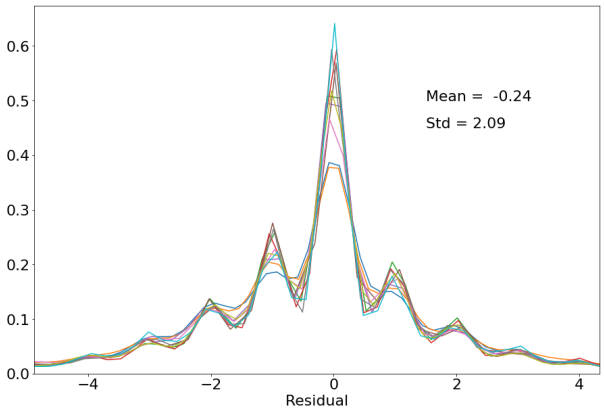


Figure 27: Residuals (Weekend) - Using information from previous steps

	Upper bound	Lower bound	Width
Weekend	-0,2327	-0,2555	0,0228
Week	-0,2853	-0,3010	0,0157

Table 18: 95% confidence interval of residuals (With trend)

6.5.4 Model for different temperatures

Section 5.3.3 showed that the average daily temperature influences the number of trips. Therefore the model has been split up into a part that is suitable for predicting 'colder' and 'hotter' days. To do so the average temperature of the days included in the dataset from 2012 until 2017 has been calculated, $22,8°C$. The dataset has then been split up into two groups, days with a average daily temperature

higher and lower than the total average temperature. These two datasets have then been separated again for week days and weekends and the model as described earlier has been set up. By setting up models for high and low temperatures it is expected that the residuals decrease, so the forecast improves.

The density plots of the residuals for week days and weekends and then separated by high and low temperatures can be seen in figure 28 and 29. The confidence intervals can be seen in table 19. Looking at the outputs it can be seen that by separating the model by the temperature the average residual was lowered. The confidence intervals during the week is also narrower but during the weekend it is wider.

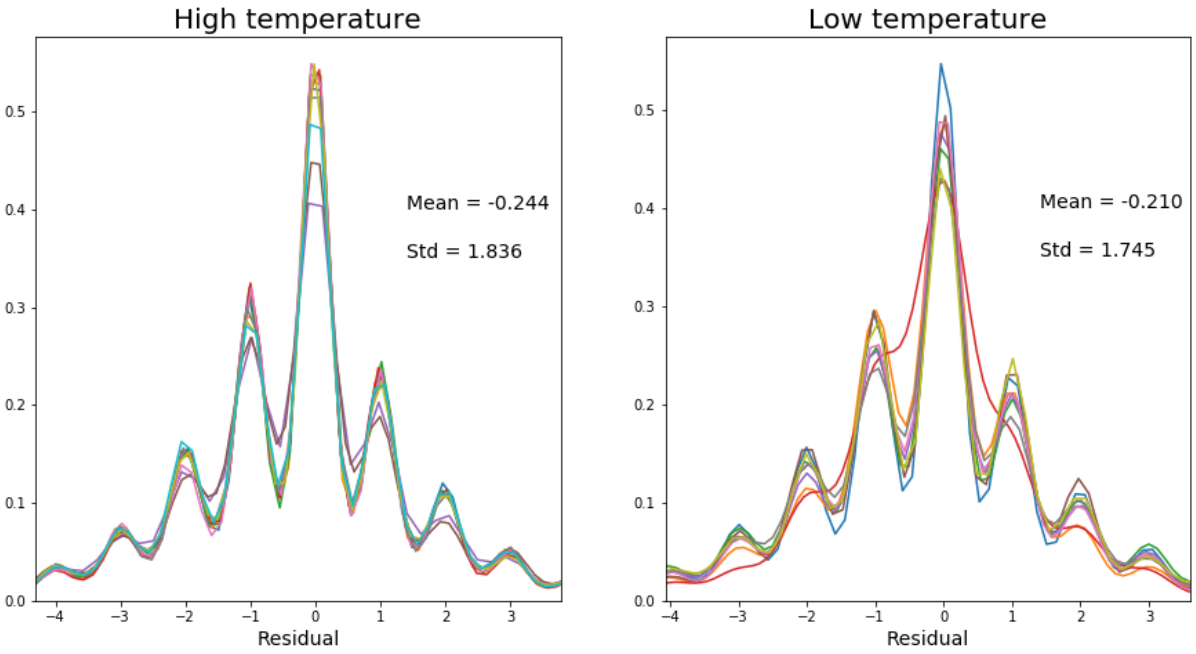


Figure 28: Residuals (Week) - Separated by temperature

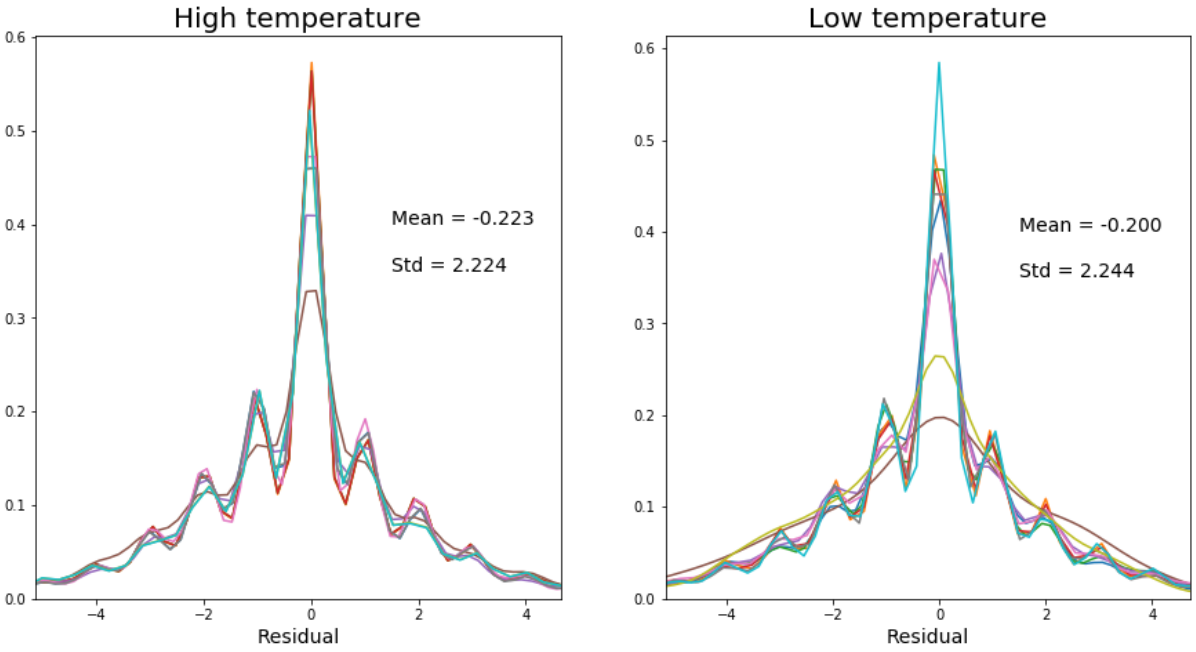


Figure 29: Residuals (Weekend) - Separated by temperature

		Upper bound	Lower bound	Width
Week	High temperature	-0,2363	-0,2515	0,0152
	Low temperature	-0,1996	-0,2205	0,0209
Weekend	High temperature	-0,2112	-0,2343	0,0231
	Low temperature	-0,1846	-0,2159	0,0313

Table 19: 95% confidence interval of residuals (With temperature)

6.6 Validation rental duration

In addition to the number of trips the model is also able to give an indication on when the bicycles will approximately arrive at the destination. In this section it is tested if the historical average mean of the rental duration is a good approximation. The outcomes for week days and weekends can be seen in figure 30 and 31. From these two figures it can be seen that the standard deviation is big, more than two hours. This is not desirable because it is unclear in which time stamp the bicycle will return.

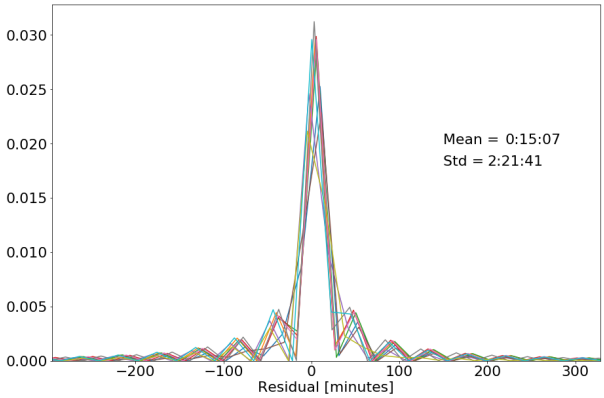


Figure 30: Residuals of rental duration (Week)

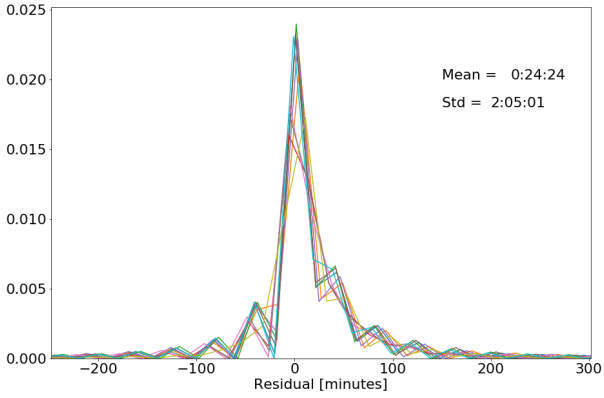


Figure 31: Residuals of rental duration (Weekend)

6.7 Summary forecasting the number of trips between OD-pairs

After identifying the factors that influence the number of trips between OD-pairs, this chapter developed a model that can forecast temporal variations in OD-pairs. For constructing the model data from 2012 until 2017 has been used because the aspect of interest was when the trips occurred. Using this information and the knowledge from chapter 5 two separate models have been created, one for forecasts during the week and one for weekends.

The basis of the two models is formed by the historical average of the number of trips per OD-pair per time stamp. This average number of trips already serves as a good estimate because the rental processes during the days and weekends are very similar. However this model does not take into account special events or situations where the demand is higher than the historical average. For this reason the difference between the observed number of trips and the predicted number of trips of the preceding time stamp is added to improve the prediction. By incorporating this information the average residual was reduced by about 50%. Also setting up models for warmer and colder days resulted in a reduction of the average residual. So the combination of historical data, information about the weather and recent information on the number of trips can forecast the number of trips between OD-pairs very accurately.

The model also includes a forecast of the time when the bikes will approximately arrive at the destination. In contrast to the number of trips the historical data is not suitable to predict estimate the rental duration. It is unknown which factors influence the rental period and therefore it is hard to estimate what would be necessary to improve the forecast.

7 Conclusions

The penultimate chapter will summarize the results from the preceded chapters and draw the conclusions of this research. To draw the conclusions it is important to recall the research objective and questions that have been defined in chapter 3. This chapter is structured as follows, first the research questions of the sub-objectives and then the general conclusions of the research are presented.

7.1 Identify the factors that influence the usage of PBSS in São Paulo

In chapter 2 various factors have been identified that influence the number of trips of a PBSS per station. Until now it was unclear if these factors also apply to the system in São Paulo. For this reason first the factors that influence trip attraction and generation at the station level have been determined.

Different simple linear regression models have been set up to identify if the factors pointed out by other researches apply to the case of São Paulo. At the station level the models that include the surrounding bicycle infrastructure, the distance to bus and CPTM stations and the job density are significant. All these models have an R-square value lower than 19% what is not very high. The highest values was observed from the model that uses the bicycle infrastructure to describe the trip attraction of a station, 0,188. Other factors that have been identified by the literature have been tested but are not significant. From the different figure in chapter 5 it can be seen that the data is very scattered what makes predicting the number of trips very difficult.

Concluding the results from the analysis that should have identified the factors that influence the number of trips per station it can be said that the factors that have been identified by the literature do not all apply to the case of São Paulo. Different personal factors that have not been investigated in this research might explain the number of trips arriving or departing from a station.

7.2 Analyse the spatial and temporal variations in the number of trips between OD-pairs of the PBSS in São Paulo

In section 5.1 the temporal and spatial patterns have been identified. Clear differences in activities between days during the week and weekends can be obtained. The total amount of trips during the week is very similar from Monday to Friday. Two clear peaks can be identified, a morning peak, ranging from approximately 07:00 to 10:00, and an evening peak, ranging from 16:00 to 20:00. In these periods most trips are made in the very west of the system. In this area the predominant land use is commercial, which means a lot of people work there. During the weekends no clear peaks can be obtained. The number of trips slowly increases, stays then at the same level for about six hours and then decreases again. On Saturday and Sunday trips often start/end close to recreational areas.

Looking at how long, time and distance wise, trips are, also differences can be observed. During the weekend the distance between origin and station is shorter than during the week. This is due to the fact that during the weekend the most frequent OD-pairs have the same origin and destination. Nonetheless the trip duration during the weekend is longer than during the week. From Monday until Friday customers are cycling more frequent between OD-pairs with longer distances. Especially during the evening rush hour long trips are made.

The factors that apply to the trip attraction and generation at the station level have been used to forecast the number of trips between OD-pairs. The analyses have been carried out for different population density areas. While investigating the relation between travel time and number of trips per OD-pair it has been discovered that in denser populated areas the travel times are shorter. This means that trips between stations that are closer are more frequent within dense populated areas. The variable that is most suitable to describe the number of trips between OD-pairs or in general system wide is the average daily temperature.

Furthermore it has been investigated if combinations of variables might be more suitable to explain the number of trips between OD-pairs. As pointed out in section 5.3.6 combining different variables to factors might be a suitable approach. Still using these factors as independent variable only explains a small portion of the variance in the number of trips per OD-pair.

As the title of this research implies a model to forecast the temporal and spatial variations has been developed. This model can give an estimate of the number of trips per OD-pair for 30-minutes intervals. Chapter 6 showed that the historical data from previous years in combination with recent information can very accurately forecast the number of trips. Separating the model for warmer and colder days even decreased the estimation error. To estimate the duration of the rental process, historical averages are not suitable.

Summarizing all the information it can be concluded that the PBSS in São Paulo is used for different purposes during the week and weekend. On the weekend the system is more used for recreational purposes, while in the week it can be assumed the system is used mostly for commuting. The temporal and spatial patterns of the rental processes are very constant and therefore good to predict. Nonetheless the factors that can be used for forecasting trip attraction and generation are not necessarily applicable to the prediction of spatial variations.

7.3 General conclusions

The main objective of this research was: *Develop a model that can forecast spatial and temporal variations in the number of trips between OD-pairs in a PBSS.* By analysing various factors this research has generated insights in the spatio-temporal patterns of the PBSS in São Paulo. The factors that have been identified in the literature do not all apply to the case of São Paulo or to OD-pairs. It must be concluded that other factors that have not been studied in this research influence the number of trips between OD-pairs and therefore the spatial and temporal variation. This factors might be socio-economic characteristics of the users. Also personal preferences with respect to cycling in general might influence the number of trips. In the case of São Paulo the aspect of security should not be neglected. Next to these factor operational issues like the amount of available or useable bikes might influence the usage. These are factors that can influence the number of trips between OD-pairs and should be investigated.

This research has shown that the number of trips between OD-pairs follow clear patterns. Clear differences between weekends and week days were obtained. These patterns are repetitive and therefore good to forecast. To forecast the temporal variations in the number of trips between OD-pairs the historical average per time stamp is a very good approximation. Nonetheless the forecast can be improved by incorporating information from the previous time stamp and the weather.

8 Discussion

The last chapter will present the discussion of the research. In this chapter the research will be critically evaluated and suggestions for further research are given. The elements of discussion will be presented step by step.

- **Used data:** For this research the data from 2017 has been used because it was available right from the start of the research. Approximately four months after the start of the research data from 2018 got available. It has been tried to incorporate the data from 2018 into the research. However from 2017 to 2018 the operator of the PBSS changed, what resulted in less stations at different location with respect to 2017. In this period the low R-square values already have been obtained. For the new locations in 2018 the relation between the number of trips and the land use around the stations has been identified. This resulted in similar low R-square values. For this reason it has been decided to stick with the 2017 because changing to the 2018 would have meant to do the analysis all over again with as outcome similar results.
- **Used data:** As mentioned in section 5.1 only the data from OD-pairs that have five or more trips are included in the research. This was an important choice because it affected which data to include and exclude from the research. In upcoming researches one might consider changing this number. This research has tested the regression models for different OD-pair groups with different number of total trips. The groups with a higher total number of trips did not have a higher R-square value than the group with more than five trips. If the number of trips would have been increased data would have been excluded, what would have meant that the results are not applicable to the whole system, only to a selected portion of the system.
- **Research approach:** The research approach that has been taken is based on data that describes the physical data concerning the trips and the environment around the stations. It is therefore hard to identify the influence of underlying factors that might explain the attitude towards cycling. As the research has shown the physical characteristics around the OD-pairs are not always suitable to describe the number of trips. Therefore it is desirable to investigate which influence socio-economic characteristics have. Also for further research it would be interesting to investigate the attitude towards cycling in general. These researches might be helpful to identify why certain OD-pairs have higher number of trips. Also these factors might identify the factors that influence the rental duration because as the research has shown the historical average is not a good estimate of the rental duration.
- **Routes between origin and destination:** The available data did not include information about which route the customers took to get from their origin to their destination. It had to be assumed that they use the shortest path. But if for example a slightly longer routes contains more bicycle infrastructure they might use the other longer route. In the research one could have dealt with this problem by identifying three possible routes from origin to destination. However due to the time limitation of the research this was not feasible.
- **Number of available bikes:** Unfortunately it was not possible to gather information about the available number of bikes per station per moment in time. It had to be assumed that at every moment in time there were enough bikes to pick up and spots available to return the bikes. The number of available bikes at the origin station influences the number of trips between an OD-pair. This is for sure a shortcoming of the research but with the available data it was inevitable.
- **Trips per bike per day:** As the research by [Gauthier et al., 2013] has shown the number of trips per bike per day can be a good measure to evaluate the performance of the public bike sharing system. Instead of using the number of arriving or departing trips to identify the factors for trip attraction and generation, one should develop a measure that combines the number of available bikes and docking points at a station and the number of trips. It is assumed that the number of available bikes and docking points is highly related to the number of trips per station. However as already mentioned for this research this was not possible due to the unavailability of data.

- **Case of São Paulo:** As mentioned in chapter 7 the factors that apply to other PBSS across the world do not all apply to the case of São Paulo. This raises the question why the system in São Paulo is so different. One could argue that the bicycle culture in South America is a different one than the one in Europe or Asia where other PBSS are located. However the work of [de Souza et al., 2017] analysed the PBSS in Rio de Janeiro and the factors that have been identified there do not apply to the case of São Paulo. During the research it was discovered that the quality of the PBSS changed many times between 2012 and 2017. Apparently in the last years the quality had decreased due to the fact that the contract with the previous operator ended and there were no incentives to ensure the quality of the service. These changes in quality might cause changes in the usage of the system.

  Another reason why the case of São Paulo is so different might be caused by the fact that the city, except for 'Parque Ibirapuera', does not have an attractive area to cycle. The PBSS are often used for recreational purposes. For this purpose people like to cycle on scenic routes where they can enjoy their trip. This can be difficult in São Paulo due to the city characteristics. As the research has shown the 'Parque Ibirapuera' is a very popular area to cycle.

- **Further research:** Next to the PBSS that use stations the recent years have seen the deployment of may so called free floating bike sharing systems. For the case of São Paulo a comparative research that identifies differences between the system would be interesting. Currently the free floating system is limited to a certain area that is similar to the operating area of the PBSS what would it make an ideal situation for comparison. Research could identify if the factors that apply to PBSS also apply to free floating systems and investigate differences in the usage patterns.

A Outputs

A.1 Additional graphs and maps

A.1.1 Additional graphs

Trip generation and attraction

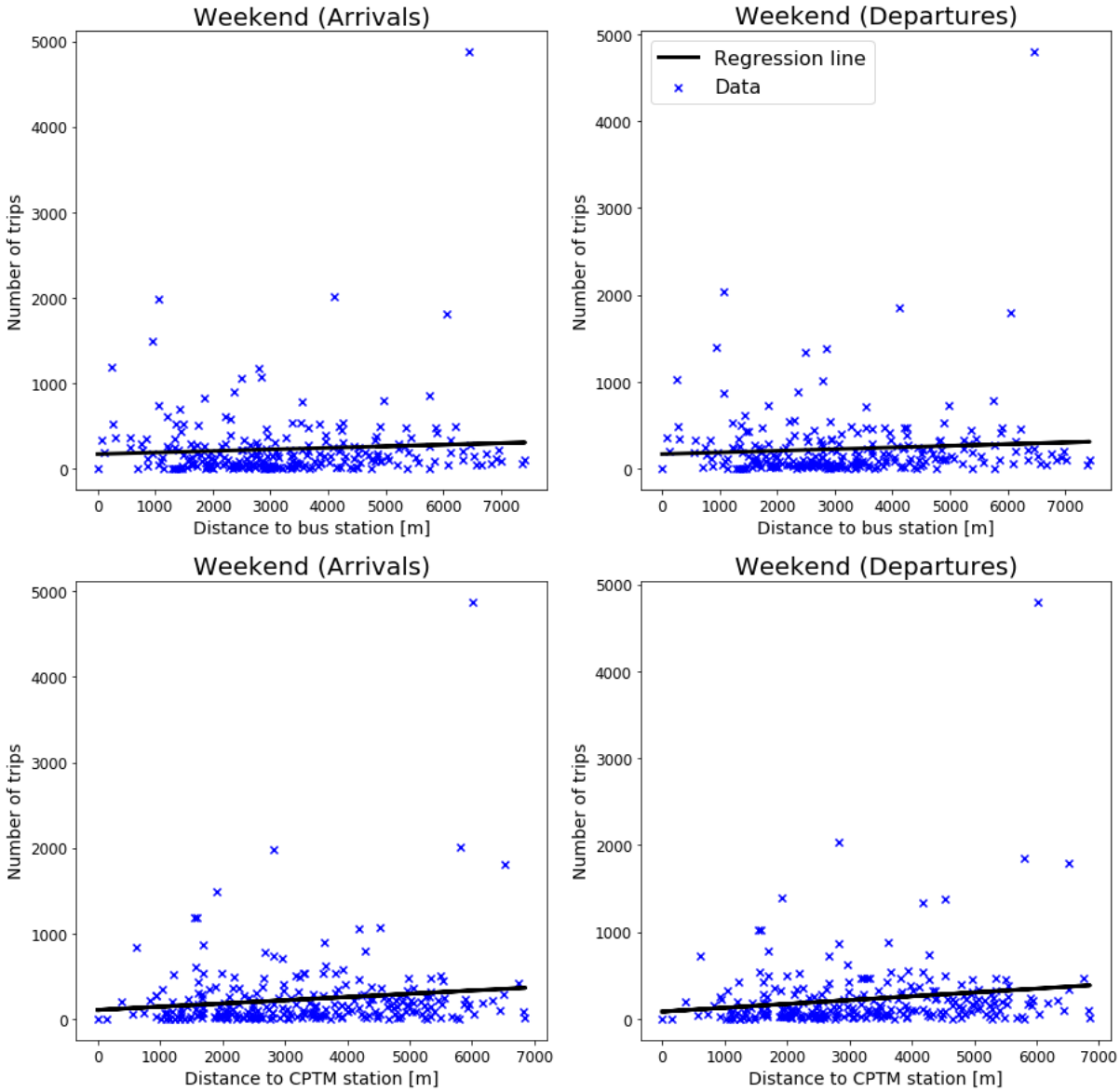


Figure 32: Relation between distance to public transport and arrivals/departures

Spatial variations

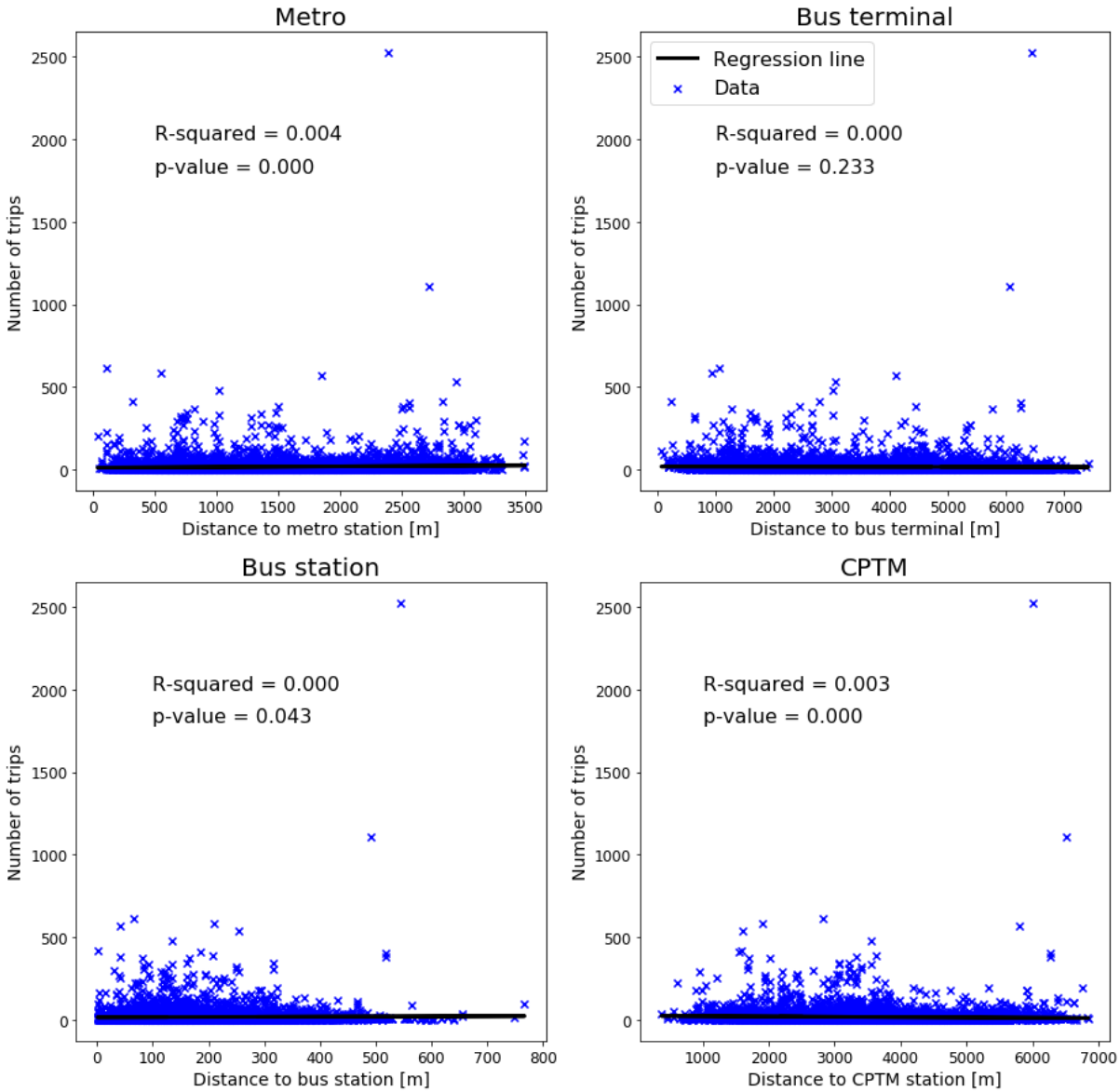


Figure 33: Regression models for distance to public transport (Week)

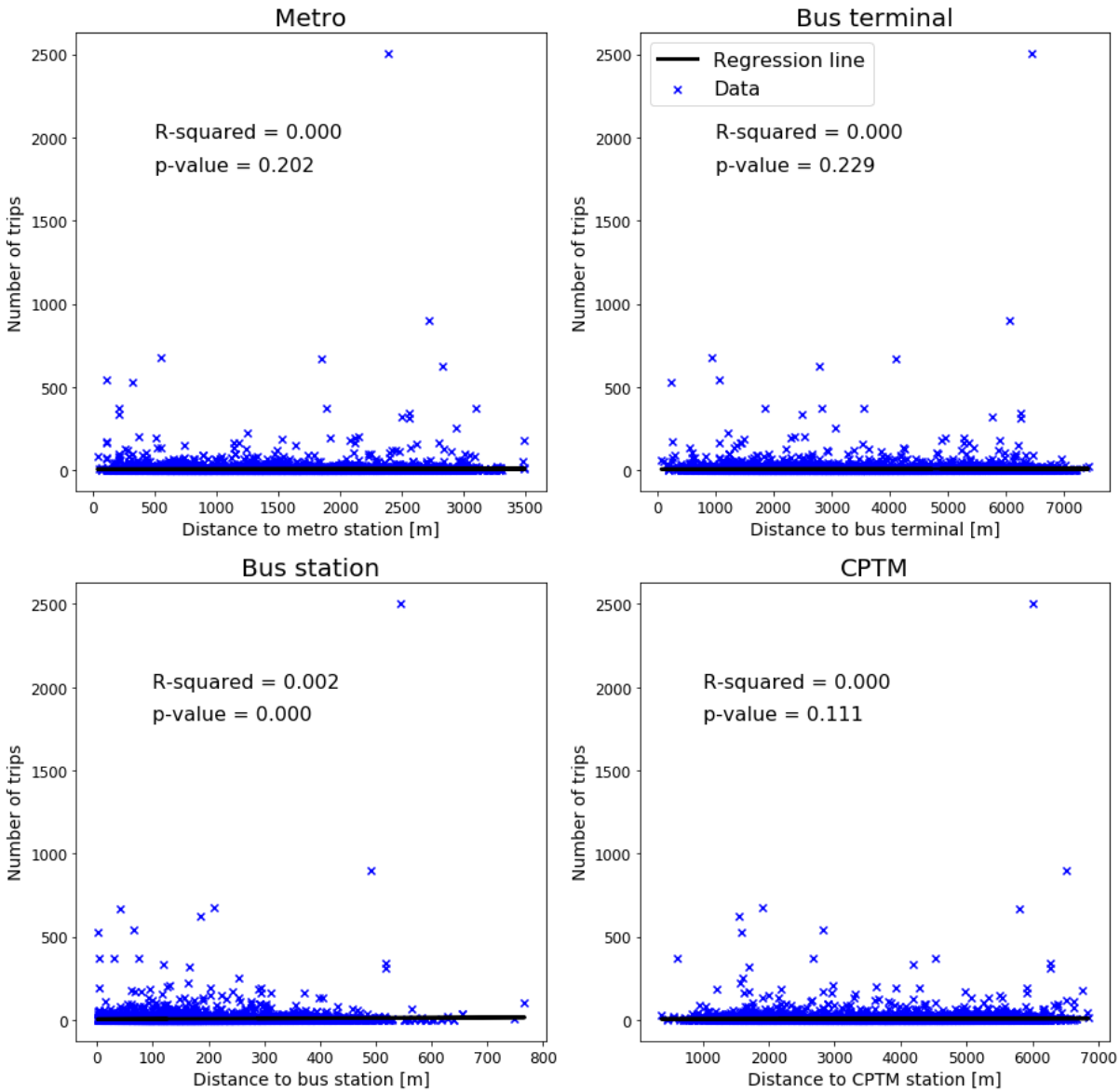


Figure 34: Regression models for distance to public transport (Weekend)

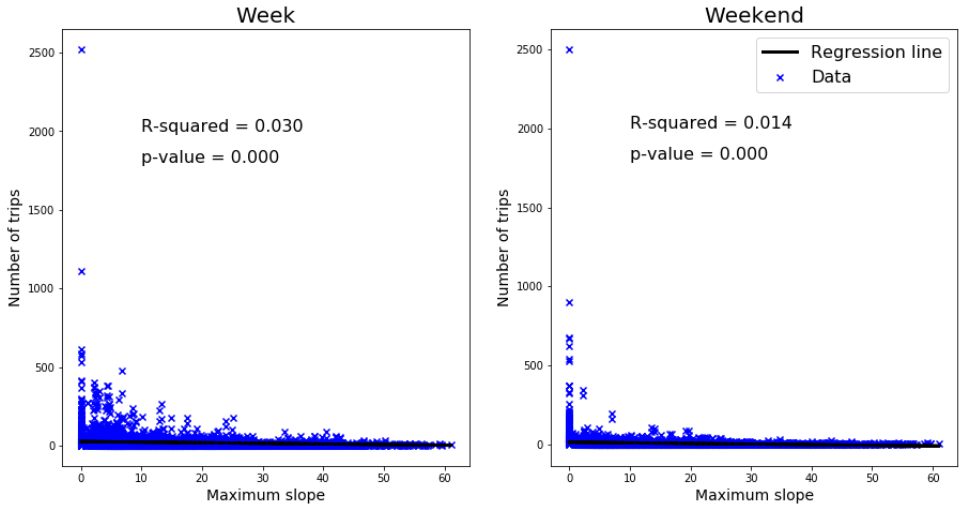


Figure 35: Regression model for maximum slope between origin and destination

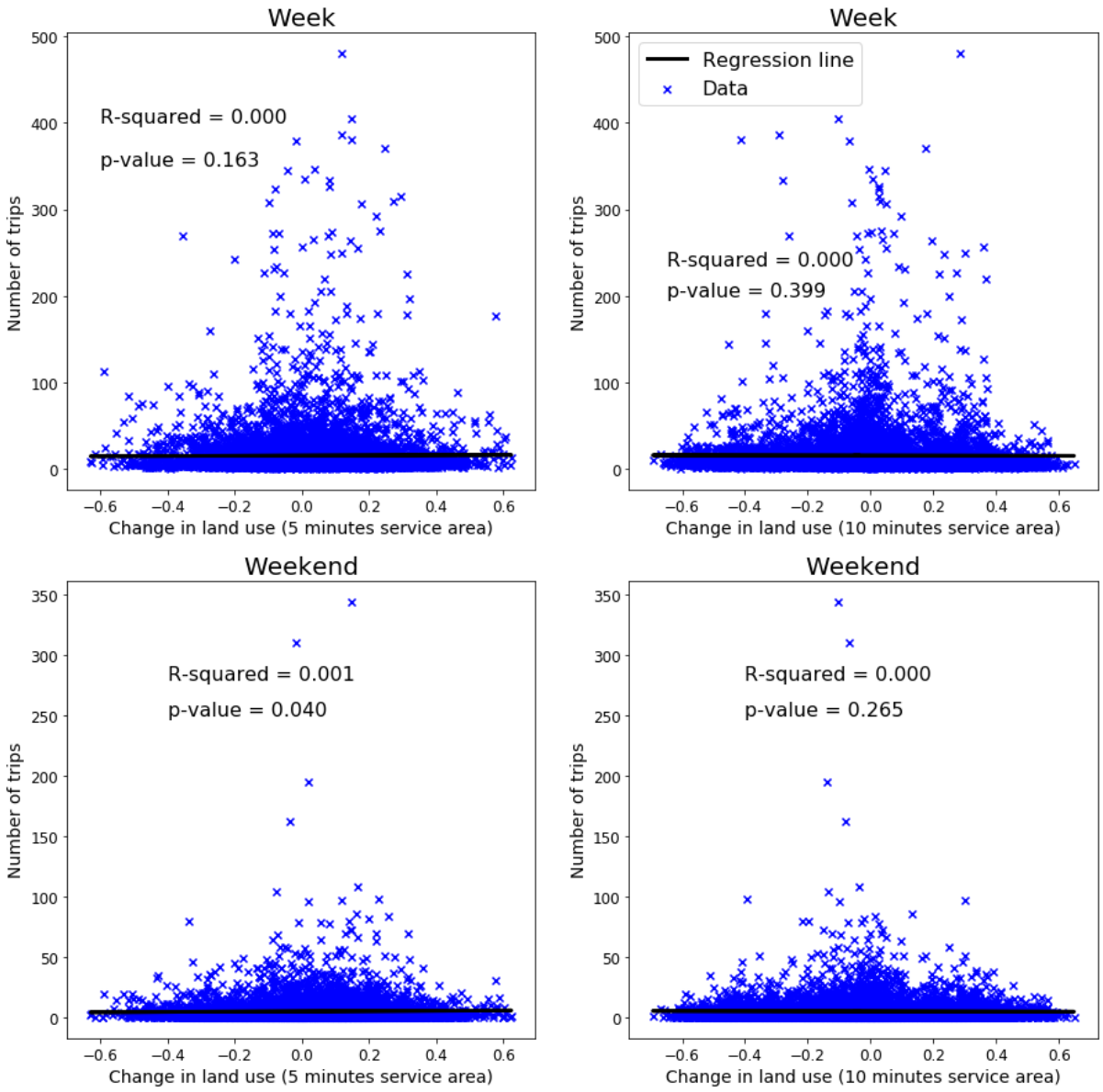


Figure 36: Regression model for change in land use between origin and destination

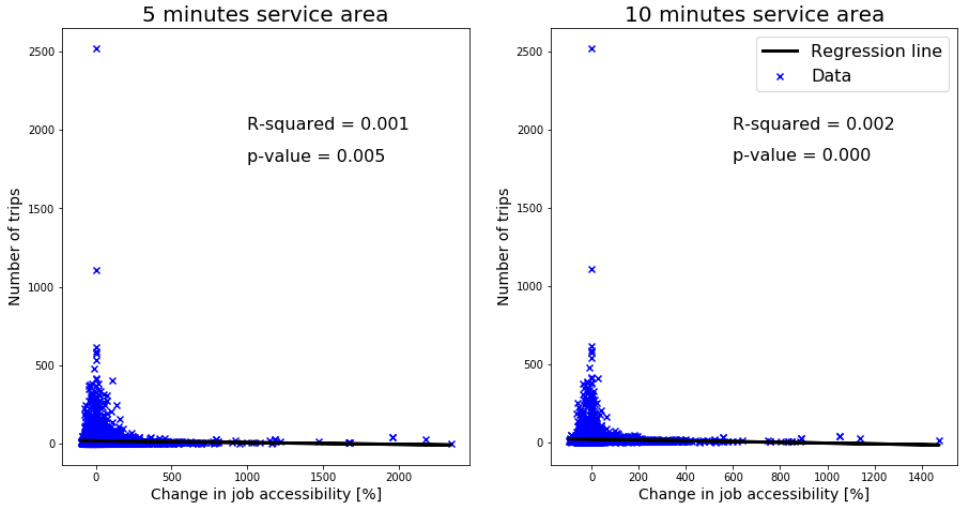


Figure 37: Change in job accessibility between origin and destination

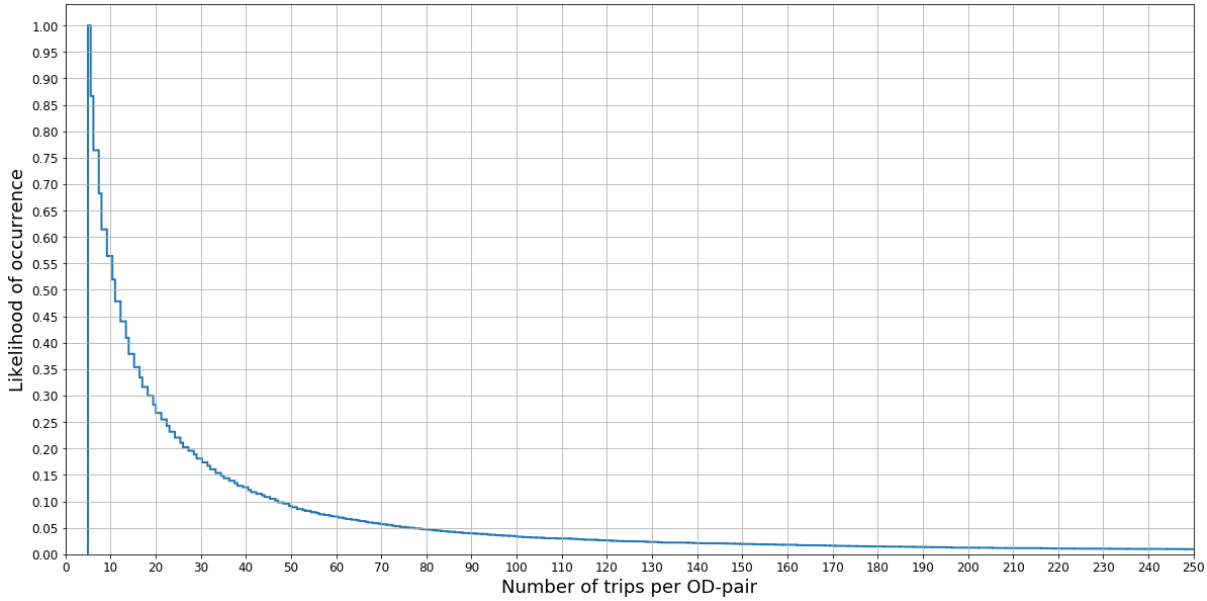


Figure 38: Cumulative density plot of the number of trips per OD-pair

A.1.2 Additional maps

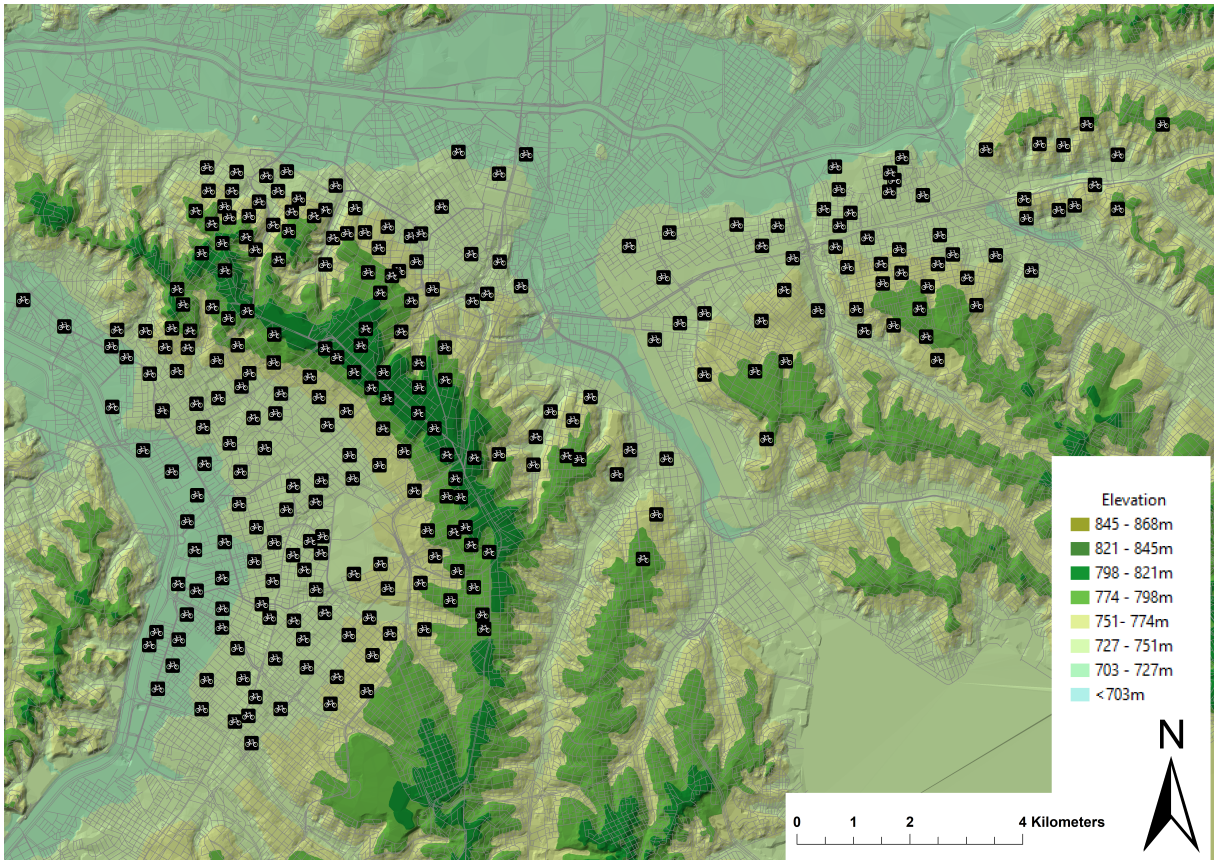


Figure 39: Topography of São Paulo

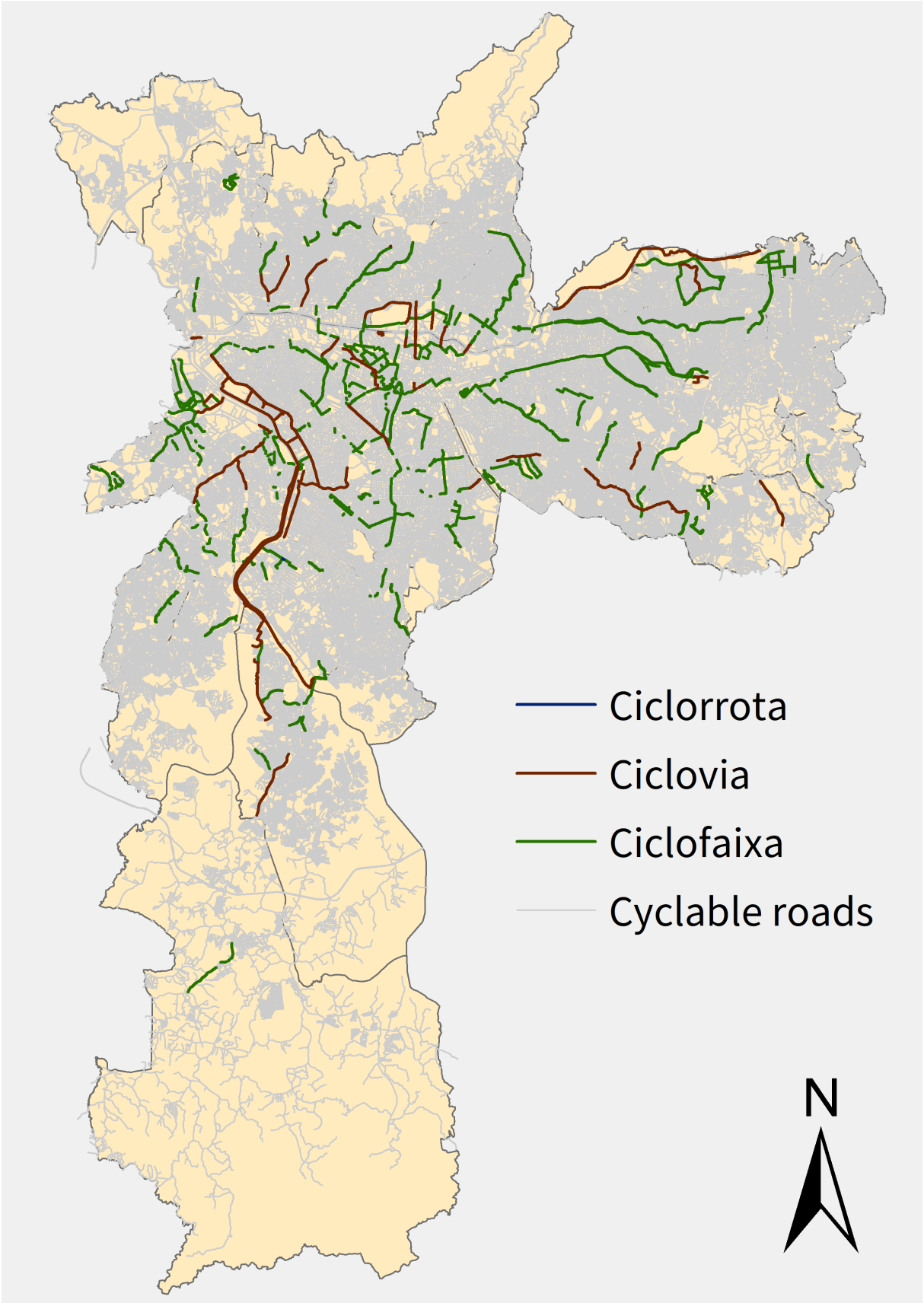


Figure 40: Bicycle infrastructure in São Paulo

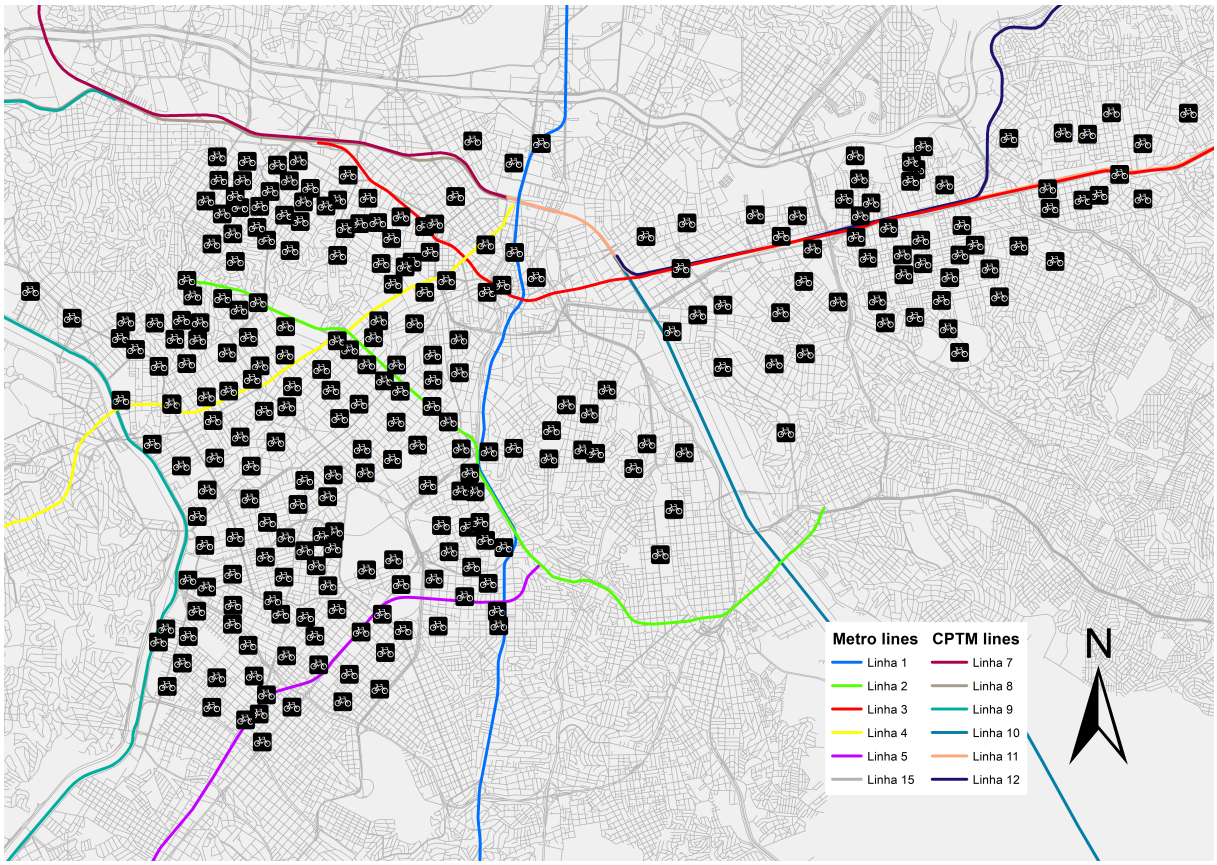


Figure 41: Metro and CPTM lines of São Paulo

A.2 Distance decay functions

A.2.1 Fitted parameters

Week			
Population density	**a**	**b**	**c**
<2160 inh/ha	1,372	0,051	-0,184
2160 - 23427 inh/ha	1,540	0,057	-0,209
23427 - 44694 inh/ha	1,550	0,053	-0,238
44694 - 65961 inh/ha	1,829	0,037	-0,535
65961 - 87228 inh/ha	1,412	0,074	-0,160
>87228 inh/ha	1,763	0,097	-0,254
Weekend			
Population density	**a**	**b**	**c**
<2160 inh/ha	1,358	0,056	-0,164
2160 - 23427 inh/ha	1,552	0,061	-0,215
23427 - 44694 inh/ha	1,525	0,058	-0,209
44694 - 65961 inh/ha	1,895	0,035	-0,624
65961 - 87228 inh/ha	1,449	0,074	-0,195
>87228 inh/ha	2,062	0,065	-0,664

Table 20: Fitted parameters for the exponential function

Week				
Population density	**Q**	**B**	**M**	**v**
<2160 inh/ha	-28,030	0,189	12,462	-7,554
2160 - 23427 inh/ha	-28,272	0,314	17,125	-10,799
23427 - 44694 inh/ha	-23,449	0,267	16,849	-9,474
44694 - 65961 inh/ha	-35,786	0,560	24,546	-21,209
65961 - 87228 inh/ha	-14,469	0,274	11,821	-7,585
>87228 inh/ha	-13,385	1,066	14,633	-21,160
Weekend				
Population density	**Q**	**B**	**M**	**v**
<2160 inh/ha	-30,847	0,242	14,674	-9,329
2160 - 23427 inh/ha	-14,690	0,342	17,772	-10,840
23427 - 44694 inh/ha	-24,167	0,271	15,026	-8,994
44694 - 65961 inh/ha	-35,361	0,595	24,180	-22,104
65961 - 87228 inh/ha	-13,789	0,320	12,738	-8,575
>87228 inh/ha	-16,215	1,263	14,473	-24,429

Table 21: Fitted parameters for the Richard's function

A.2.2 Residuals of distance decay functions

Week		
Population density	**Exponential function**	**Richard's function**
<2160 inh/ha	0,02	0,01
2160 - 23427 inh/ha	0,04	0,02
23427 - 44694 inh/ha	0,04	0,03
44694 - 65961 inh/ha	0,04	0,03
65961 - 87228 inh/ha	0,03	0,02
>87228 inh/ha	0,04	0,02

Weekend		
Population density	**Exponential function**	**Richard's function**
<2160 inh/ha	0,03	0,01
2160 - 23427 inh/ha	0,04	0,02
23427 - 44694 inh/ha	0,04	0,03
44694 - 65961 inh/ha	0,04	0,03
65961 - 87228 inh/ha	0,03	0,02
>87228 inh/ha	0,04	0,02

Table 22: RMSE per fit (Week & Weekend)

A.3 Likelihood of change in job accessibility

A.3.1 Fitted parameters

Population density	**Q**	**B**	**M**	**v**
<2160 inh/ha	-18,620	0,036	10,612	-8,672
2160 - 23427 inh/ha	-13,168	0,038	10,742	-7,980
23427 - 44694 inh/ha	-3,824	0,018	6,792	-4,223
44694 - 65961 inh/ha	0,301	0,005	-4,893	0,867
65961 - 87228 inh/ha	-16,699	0,027	4,623	-6,470
>87228 inh/ha	-49,999	0,053	49,999	-13,438

Table 23: Fitted parameters for the Richard's function (5 minutes service area)

Population density	Q	B	M	v
<2160 inh/ha	-42,658	0,107	16,329	-17,711
2160 - 23427 inh/ha	-19,206	0,099	25,032	-15,989
23427 - 44694 inh/ha	-3,956	0,023	11,189	-4,592
44694 - 65961 inh/ha	-1,151	0,010	-4,999	-1,736
65961 - 87228 inh/ha	-49,999	0,064	49,999	-14,556
>87228 inh/ha	-49,999	0,080	49,999	-16,101

Table 24: Fitted parameters for the Richard's function (10 minutes service area)

A.3.2 Residuals of change in job accessibility

Population density	5 minutes service area	10 minutes service area
<2160 inh/ha	0,03	0,06
2160 - 23427 inh/ha	0,04	0,05
23427 - 44694 inh/ha	0,02	0,03
44694 - 65961 inh/ha	0,01	0,02
65961 - 87228 inh/ha	0,04	0,06
>87228 inh/ha	0,04	0,06

Table 25: RMSE for Richard's function

A.4 Parameters of regression models (Weather)

Dependent variable	Constant	Coefficient of temperature
Number of trips with different origin and destination (Week)	-816,035	66,952
Number of trips with different origin and destination (Weekend)	-747,712	57,640
Number of trips with same origin and destination (Week)	-123,999	10,736
Number of trips with same origin and destination (Weekend)	-393,935	29,750

Table 26: Regression model weather (Estimated coefficients)

A.5 Factor analysis

	Loadings					
Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Change Land use (5 minutes)	-0,040	-0,043	0,055	0,058	-0,014	**0,949**
Change Land use (10 minutes)	-0,098	0,151	**-0,387**	-0,295	0,232	0,296
Dist to Metro	-0,051	0,166	0,401	**-0,469**	0,517	-0,144
Dist to bus terminal	0,027	0,091	**0,861**	-0,272	0,196	0,020
Dist to bus station	-0,011	**0,988**	0,101	-0,059	0,009	-0,012
Dist to CPTM	-0,007	0,184	**0,837**	0,045	-0,204	0,084
Dist to all PT	-0,011	**0,986**	0,113	-0,068	0,014	-0,014
Travel time	-0,226	0,016	0,109	**0,711**	0,402	-0,024
BnR Accessibility (5 minutes)	**0,955**	0,003	0,008	-0,056	0,040	-0,043
BnR Accessibility (10 minutes)	**0,964**	-0,030	0,023	-0,034	-0,024	-0,015
PT Accessibility (5 minutes)	**0,965**	0,009	0,005	-0,065	0,028	-0,028
PT Accessibility (10 minutes)	**-0,961**	0,019	-0,017	0,051	0,030	0,004
Maximum slope	-0,047	-0,085	-0,164	**0,829**	-0,029	0,044
Share of bicycle paths	0,064	-0,029	-0,119	0,150	**0,771**	0,038

Table 27: Load factors for variables (Weekend)

Variables	Loadings					
	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
Change Land use (5 minutes)	-0,066	-0,094	0,107	0,113	-0,002	**0,902**
Change Land use (10 minutes)	-0,075	0,242	-0,409	-0,258	0,117	**0,414**
Dist to Metro	-0,047	0,222	0,365	**-0,651**	0,337	-0,121
Dist to bus terminal	0,036	0,194	**0,840**	-0,341	0,098	0,025
Dist to bus station	-0,023	**0,973**	0,147	-0,052	-0,001	-0,022
Dist to CPTM	0,019	0,203	**0,842**	0,096	-0,154	0,085
Dist to all PT	-0,024	**0,971**	0,157	-0,061	0,001	-0,021
Travel time	-0,099	0,130	0,172	0,556	**0,590**	-0,010
BnR Accessibility (5 minutes)	**0,936**	-0,012	-0,003	-0,067	0,049	-0,101
BnR Accessibility (10 minutes)	**0,953**	-0,034	0,042	0,038	-0,038	0,020
PT Accessibility (5 minutes)	**0,949**	0,000	-0,001	-0,063	0,038	-0,088
PT Accessibility (10 minutes)	**-0,947**	0,022	-0,035	-0,033	0,049	-0,033
Maximum slope	-0,036	-0,038	-0,009	**0,831**	0,148	0,011
Share of bicycle paths	0,047	-0,054	-0,141	-0,008	**0,812**	0,039

Table 28: Load factors for variables (Weekend)

A.5.1 T-test factor analysis

Factor		Lowest 50 OD-pairs	Highest 50 OD-pairs
1	Mean	11,18	11,84
	Stdev	9,500926	8,635647
	p-value	0,719742	
2	Mean	36	101,6
	Stdev	62,60895	384,9946
	p-value	0,244487	
3	Mean	48,66	48,78
	Stdev	64,68032	156,1726
	p-value	0,99605	
4	Mean	61,44	42,26
	Stdev	98,23343	54,61897
	p-value	0,235958	
5	Mean	32,7	55,9
	Stdev	34,60881	63,20799
	p-value	0,027106	
6	Mean	19,98	22,14
	Stdev	23,271	27,14554
	p-value	0,673332	

Table 29: T-test results (Week)

Factor		Lowest 50 OD-pairs	Highest 50 OD-pairs
1	Mean	12,32	9,58
	Stdev	12,2938	5,713458
	p-value	0,160294	
2	Mean	40,14	96,72
	Stdev	90,66995	370,0053
	p-value	0,30306	
3	Mean	33,38	106,3
	Stdev	75,80261	369,002
	p-value	0,18115	
4	Mean	54,7	28,88
	Stdev	101,1611	57,82928
	p-value	0,124929	
5	Mean	50,92	7,5
	Stdev	99,38286	3,41321
	p-value	0,003617	
6	Mean	11,26	8,52
	Stdev	8,449402	4,838347
	p-value	0,051675	

Table 30: T-test results (Weekend)

Bibliography

[Askham, 2013] Askham, N. (2013). DEFINING DATA QUALITY DIMENSIONS. Technical report, DAMA UK Working Group.

[Borgnat et al., 2011] Borgnat, P., Abry, P., Flandrin, P., Robardet, C., Rouquier, J.-B., and Fleury, E. (2011). Shared Bicycles in a City: a Signal Processing and Data Analysis Perspective. *Advances in Complex Systems*, 14(3):415–438.

[Broach et al., 2012] Broach, J., Dill, J., and Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, 46(10):1730–1740.

[Cleland and Walton, 2004] Cleland, B. S. and Walton, D. (2004). Why don't people walk and cycle? Technical report, Lower Hutt.

[Côme et al., 2014] Côme, E., Randriamanamihaga, N. A., Oukhellou, L., and Aknin, P. (2014). Spatio-temporal analysis of Dynamic Origin-Destination data using Latent Dirichlet Allocation: Application to the Vélib ' Bike Sharing System of Paris . *TRB 93rd Annual meeting*, pages 1–18.

[Corcoran et al., 2014] Corcoran, J., Li, T., Rohde, D., Charles-Edwards, E., and Mateo-Babiano, D. (2014). Spatio-temporal patterns of a Public Bicycle Sharing Program: the effect of weather and calendar events. *Journal of Transport Geography*, 41:292–305.

[de Souza et al., 2017] de Souza, F., La Paix Puello, L., Brussel, M., Orrico, R., and van Maarseveen, M. (2017). Modelling the potential for cycling in access trips to bus, train and metro in Rio de Janeiro. *Transportation Research Part D: Transport and Environment*, 56:55–67.

[Demaio, 2009] Demaio, P. (2009). Bike-sharing : History , Impacts , Models of Provision , and Future. *Journal of Public Transportation*, 12(DeMaio 2004):41–56.

[Etienne and Latifa, 2014] Etienne, C. and Latifa, O. (2014). Model-Based Count Series Clustering for Bike Sharing System Usage Mining: A Case Study with the Vélib' System of Paris. *ACM Trans. Intell. Syst. Technol*, 5(3).

[Fecchio, 2018] Fecchio, L. G. (2018). Fatores que influenciam o uso da bicicleta no acesso e integração com o metrô: Estudo de caso na linha 4-Amarela do metrô de São Paulo. Technical report, Escola Politécnica da Universidade de São Paulo, São Paulo.

[Feng et al., 2017] Feng, Y., Affonso, R. C., and Zolghadri, M. (2017). Analysis of bike sharing system by clustering: the Vélib' case. *IFAC-PapersOnLine*, 50(1):12422–12427.

[Frank et al., 2006] Frank, L. D., Sallis, J. F., Conway, T. L., Chapman, J. E., Saelens, B. E., and Bachman, W. (2006). Many Pathways from Land Use to Health Associations between Neighborhood Walkability and Active Transportation, Body Mass Index, and Air Quality. Technical report.

[Freitas and Maciel, 2017a] Freitas, A. L. P. and Maciel, A. B. L. (2017a). Assessing Cyclists' Perceptions, Motivations and Behaviors: An Exploratory Study in Brazil. In *Procedia Engineering*, volume 198, pages 26–33, Shanghai. Elsevier.

[Freitas and Maciel, 2017b] Freitas, A. L. P. and Maciel, A. B. L. (2017b). Cycling in a Brazilian City. In *Procedia Engineering*, volume 198, pages 411–418, Shanghai. Elsevier.

[Froehlich et al., 2009] Froehlich, J., Neumann, J., and Oliver, N. (2009). Sensing and Predicting the Pulse of the City through Shared Bicycling. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, pages 1420–1426, Pasadena, California, USA. Morgan Kaufmann Publishers Inc.

[Fuller et al., 2011] Fuller, D., Gauvin, L., Kestens, Y., Daniel, M., Fournier, M., Morency, P., and Drouin, L. (2011). Use of a New Public Bicycle Share Program in Montreal, Canada. *American Journal of Preventive Medicine*, 41(1):80–83.

[Gast et al., 2015] Gast, N., Massonnet, G., Reijsbergen, D., and Tribastone, M. (2015). Probabilistic Forecasts of Bike-Sharing Systems for Journey Planning. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, pages 703–712.

[Gauthier et al., 2013] Gauthier, A., Colin, H., Kost, C., Li, S., Linke, C., Lotshaw, S., Mason, J., Pardo, C., Rasore, C., Schroeder, B., Treviño, X., Van Eyken, C., Magnusson, J., and Lewenstein, G. (2013). The Bike-Sharing Planning Guide. Technical report, New-York.

[González et al., 2016] González, F., Melo-Riquelme, C., and de Grange, L. (2016). A combined destination and route choice model for a bicycle sharing system. *Transportation*, 43(3):407–423.

[Guenther and Bradley, 2013] Guenther, M. C. and Bradley, J. T. (2013). Journey Data Based Arrival Forecasting for Bicycle Hire Schemes. In *Analytical and Stochastic Modelling Techniques and Applications*, volume 7984, pages 214–231, Ghent.

[Hui et al., 2010] Hui, Z., Lei, Y. U., Jifu, G., Nale, Z., Huimin, W., and Lin, Z. (2010). Estimation of Time-Varying OD Demands Incorporating FCD and RTMS Data. *Journal of Transportation Systems Engineering and Information Technology*, 10(1):72–80.

[Hyland et al., 2018] Hyland, M., Hong, Z., Pinto, H. K. R. d. F., and Chen, Y. (2018). Hybrid cluster-regression approach to model bikeshare station usage. *Transportation Research Part A: Policy and Practice*, 115:71–89.

[Immers and Stada, 1998] Immers, L. H. and Stada, J. E. (1998). Traffic Demand Modelling. Technical report, Katholieke Universiteit Leuven, Leuven.

[Ji et al., 2014] Ji, S., Cherry, C. R., Han, L. D., and Jordan, D. A. (2014). Electric bike sharing: Simulation of user demand and system availability. *Journal of Cleaner Production*, 85:250–257.

[Kaltenbrunner et al., 2010] Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., and Banchs, R. (2010). Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466.

[Kim, 2018] Kim, K. (2018). Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations. *Journal of Transport Geography*, 66:309–320.

[Kou and Cai, 2019] Kou, Z. and Cai, H. (2019). Understanding bike sharing travel patterns: An analysis of trip data from eight cities. *Physica A: Statistical Mechanics and its Applications*, 515:785–797.

[Krykewycz et al., 2010] Krykewycz, G. R., Puchalsky, C. M., Rocks, J., Bonnette, B., and Jaskiewicz, F. (2010). Defining a Primary Market and Estimating Demand for Major Bicycle-Sharing Program in Philadelphia, Pennsylvania. *Transportation Research Record: Journal of the Transportation Research Board*, 2143:117–124.

[Lathia et al., 2012] Lathia, N., Ahmed, S., and Capra, L. (2012). Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies*, 22:88–102.

[Li et al., 2015] Li, Y., Zheng, Y., Zhang, H., and Chen, L. (2015). Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15*, pages 1–10, Bellevue.

[Lu et al., 2013] Lu, C.-C., Zhou, X., and Zhang, K. (2013). Dynamic origin-destination demand flow estimation under congested traffic conditions. *Transportation Research Part C*, 34:16–37.

[Ma and Qian, 2018] Ma, W. and Qian, Z. S. (2018). Estimating multi-year 24 / 7 origin-destination demand using high-granular multi-source traffic data. *Transportation Research Part C: Emerging Technologies*, 96:96–121.

[Martens, 2004] Martens, K. (2004). The bicycle as a feedering mode: experiences from three European countries. *Transportation Research Part D: Transport and Environment*, 9(4):281–294.

[Martínez and Viegas, 2013] Martínez, L. M. and Viegas, J. M. (2013). A new approach to modelling distance-decay functions for accessibility assessment in transport studies. *Journal of Transport Geography*, 26:87–96.

[Mateo-Babiano et al., 2016] Mateo-Babiano, I., Bean, R., Corcoran, J., and Pojani, D. (2016). How does our natural and built environment affect the use of bicycle sharing? *Transportation Research Part A: Policy and Practice*, 94.

[McBain and Caulfield, 2017] McBain, C. and Caulfield, B. (2017). An analysis of the factors influencing journey time variation in the cork public bike system. *Sustainable Cities and Society*.

[McBain and Caulfield, 2018] McBain, C. and Caulfield, B. (2018). An analysis of the factors influencing journey time variation in the cork public bike system. *Sustainable Cities and Society*, 42:641–649.

[McLoughlin et al., 2012] McLoughlin, I. V., Narendra, I. K., Koh, L. H., Nguyen, Q. H., Seshadri, B., Zeng, W., and Yao, C. (2012). Campus Mobility for the Future: The Electric Bicycle. *Journal of Transportation Technologies*, 02(01):1–12.

[Médard de Chardon, 2016] Médard de Chardon, C. (2016). *A GEOGRAPHICAL ANALYSIS OF BICYCLE SHARING SYSTEMS*. PhD thesis, UNIVERSITÉ DU LUXEMBOURG.

[Menon et al., 2015] Menon, A. K., Cai, C., Wang, W., Wen, T., and Chen, F. (2015). Fine-grained OD estimation with automated zoning and sparsity regularisation. *Transportation Research Part B*, 80:150–172.

[Nielsen and Skov-Petersen, 2018] Nielsen, T. A. S. and Skov-Petersen, H. (2018). Bikeability – Urban structures supporting cycling. Effects of local, urban and regional scale urban form factors on cycling from home and workplace locations in Denmark. *Journal of Transport Geography*, 69:36–44.

[Peterson, 2007] Peterson, A. (2007). *The Origin-Destination Matrix Estimation Problem-Analysis and Computations*. PhD thesis, Linköping University.

[Polat, 2012] Polat, C. (2012). The Demand Determinants for Urban Public Transport Services: A Review of the Literature. *Journal of Applied Science*, 12(12):1211–1231.

[Pritchard et al., 2019] Pritchard, J. P., Tomasiello, D. B., Giannotti, M., and Geurs, K. (2019). Potential impacts of bike-and-ride on job accessibility and spatial equity in São Paulo, Brazil. *Transportation Research Part A: Policy and Practice*.

[Rahn, 2018] Rahn, M. (2018). Factor Analysis: A Short Introduction, Part 1 - The Analysis Factor.

[Raux et al., 2017] Raux, C., Zoubir, A., and Geyik, M. (2017). Who are bike sharing schemes members and do they travel differently? The case of Lyon's "Velo'v" scheme. *Transportation Research Part A: Policy and Practice*, 106:350–363.

[Rixey, 2013] Rixey, R. A. (2013). Station-Level Forecasting of Bikesharing Ridership. *Transportation Research Record: Journal of the Transportation Research Board*, 2387(2387):46–55.

[Rixey and Ranaiefar, 2016] Rixey, R. A. and Ranaiefar, F. (2016). TRB 2016 Annual Meeting Original paper submittal - not revised by author. Technical report.

[Seskin et al., 1996] Seskin, S., Cervero, R., and Zupan, J. (1996). *Transit and Urban Form*, volume 1. Portland.

[Shaheen et al., 2010] Shaheen, S., Guzman, S., and Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia. *Transportation Research Record: Journal of the Transportation Research Board*, 2143:159–167.

[Shelat et al., 2018] Shelat, S., Huisman, R., and van Oort, N. (2018). Analysing the trip and user characteristics of the combined bicycle and transit mode. *Research in Transportation Economics*.

[Singhvi et al., 2015] Singhvi, D., Singhvi, S., Frazier, P. I., Henderson, S. G., Mahony, E. O., Shmoys, D. B., and Woodard, D. B. (2015). Predicting Bike Usage for New York City ' s Bike Sharing System. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 110–114.

[Tsekeris and Tsekeris, 2015] Tsekeris, T. and Tsekeris, C. (2015). DEMAND FORECASTING IN TRANSPORT: OVERVIEW AND MODELING ADVANCES. *Economic Research-Ekonomska Istraživanja*, 24(1):82–94.

[Vogel and Mattfeld, 2011] Vogel, P. and Mattfeld, D. C. (2011). Strategic and Operational Planning of Bike-Sharing Systems by Data Mining - A Case Study. In *Computational Logistics*, pages 127–141, Hamburg. Springer-Verlag Berlin Heidelberg.

[Wang and Lindsey, 2019] Wang, J. and Lindsey, G. (2019). Do new bike share stations increase member use : A quasi- experimental study. *Transportation Research Part A*, 121(September 2018):1–11.

[Wang et al., 2016] Wang, J., Tsai, C.-H., and Lin, P.-C. (2016). Applying spatial-temporal analysis and retail location theory to public bikes site selection in Taipei. *Transportation Research Part A: Policy and Practice*, 94:45–61.

[Xu et al., 2018] Xu, C., Ji, J., and Liu, P. (2018). The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation Research Part C: Emerging Technologies*, 95:47–60.

[Yang et al., 2016] Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., and Moscibroda, T. (2016). Mobility Modeling and Prediction in Bike-Sharing Systems. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '16*, pages 165–178, Singapore.

[Yoon et al., 2012] Yoon, J. W., Pinelli, F., and Calabrese, F. (2012). Cityride: A predictive bike sharing journey advisor. In *2012 IEEE 13th International Conference on Mobile Data Management, MDM 2012*, pages 306–311.

[Zhang et al., 2017] Zhang, D., Yu, C., Desai, J., Lau, H., and Srivathsan, S. (2017). A time-space network flow approach to dynamic repositioning in bicycle sharing systems. *Transportation Research Part B: Methodological*, 103:188–207.

[Zhang et al., 2018] Zhang, Y., Brussel, M., Thomas, T., and van Maarseveen, M. (2018). Mining bike-sharing travel behavior data: An investigation into trip chains and transition activities. *Computers, Environment and Urban Systems*, 69:39–50.

[Zhao and Li, 2017] Zhao, P. and Li, S. (2017). Bicycle-metro integration in a growing city: The determinants of cycling as a transfer mode in metro station areas in Beijing. *Transportation Research Part A: Policy and Practice*, 99:46–60.

[Zhou and Mahmassani, 2007] Zhou, X. and Mahmassani, H. S. (2007). A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. *Transportation Research Part B: Methodological*, 41(8):823–840.