University of Tartu
Institute of Computer Science
Innovation and Technology Management Curriculum

Taron Davtyan

Forecasting Bicycle Demand: Bologna Case Study

Master's Thesis

Supervisors:

Rajesh Sharma PhD
Flavio Bertini PhD

Tartu 2020

Forecasting Bicycle Demand: Bologna Case Study

Abstract:

Although there are a large number of academical studies conducted about demand forecasting in docked bike-sharing programs, there is scarce literature on the dockless bike-sharing programs and especially in forecasting demand using a deep learning approach. Dockless bike-sharing programs have been growing rapidly during the past few years and having a model that can accurately predict bike usage is becoming essential for bike-sharing companies and governmental institutions. This research paper aims to develop a model to forecast the usage of private bicycles with a deep learning approach and fill the research gap mentioned above. For predicting the number of rides, long short-term memory (LSTM) neural networks model was developed. The model was used to predict bike usage for 30-minute and 60-minute intervals. Besides the historical number usage of bikes, the prediction model considers air temperature, precipitation amount, and national holidays. The study results suggest that prediction with the LSTM model gives a more accurate outcome than more widely used machine learning algorithms such as linear regression, Random Forrest, and XGBoost. LSTM model that was developed by this study can be used to predict the utilization of bike lanes, which can be essential for governmental institutions and can also help bike-sharing companies to distribute bikes across the city to provide more convenient experience to the users.

CERCS: P170 Computer science, numerical analysis, systems, control

Keywords: dockless bike-sharing, Spatial analysis, Demand forecasting, Neural networks, Long-short term memory,

Jalgratta nõudluse prognoosimine: Bologna juhtumianalüüs

Tänase päeva seisuga on tehtud palju akadeemilisi töid, mis uurivad prognoositavat nõudlust dokkides jalgrataste jagamise programmidele, kuid on tehtud vähe töid, on spetsialiseeritud dokkideta rataste jagamise kohta. Spetsialiseerumisele lisaks, on vähesed tööd uurinud nende nõudlust kasutades sügavõppe lähenemist. Dokkideta rataste jagamise populaarsus on viimaste aastate jooksul näinud kiiret kasvu, mis tõstab vajadust mudeli järgi, mis suudab täpselt prognoosida rataste kasutamist. Sellise mudeli olemasolu on eriti hädavajalik ettevõtetele ja valitsuste institutsioonidele, kes tegelevad rataste jagamisega. Eelmainitud probleemi lahendamiseks on antud uurimustöö eesmärgiks välja arendada mudel, mis suudab prognoosida dokkideta rataste kasutust sügavõppe meetodil. Selleks, et prognoosida rattasõitude arvu, arendati uurimistöös välja long short-term memory (LSTM) närvivõrgustiku mudel. LSTM mudelit kasutati, et välja uurida rataste kasutamise hulk 30-ja 60-minutiliste intervallidega. Lisaks mineviku rattasõitude arvule,

võtab mudel arvesse ka õhu temperatuuri, sademete hulga ning riigipühad. Uurimistöö tulemustest võib järeldada, et LSTM mudel annab parema resultaadi, kui palju kasutatud masinõppe algoritmid, milleks on populaarsed lineaarregressioonid, Random Forrest, ja XGBoost. Uurimistöös välja arendatud LSTM mudeliga saab prognoosida jalgrataste radade utiliseerimist, mis võib olla hädavajalik valitsuse institutsioonidele. Samuti võivad seda kasutada jalgrattaid jagavad ettevõtted, uurides jalgrataste jagamise viise linna sees, mis võib efektiivse tulemuse korral viia kasutajas mugavuste suurenemiseni.

CERCS: P170 Arvutiteadus, arvanalüüs, süsteemid, kontroll

Contents

1 Introduction

In many countries, cycling is one of the most popular types of transportation. It is constantly being promoted by local governments to be used as the main method of transportation. The main reason for such promotions is that bikes are environment-friendly and affordable, also cycling lightens road traffic, and overall it is a healthy activity. Nowadays, many countries are keen to create more bike lanes and popularize the usage of cycling. But the major problem is the utilization of those lanes for better traffic management, something that some countries are not able to predict accurately and are left with the same busy roads and deserted bike lanes.

In recent years several studies were conducted regarding bike-sharing services. Many authors analyzed what are the main factors contributing to the number of rides done by bikes. Some of them focused their studies on effects of weather and precipitation [1], [2], [3]. Others tried to understand the patterns of bike rides during different times of the days, months and seasons [4], [5]. Many researchers focused on analyzing spatial patterns of bike usage in the cities [6], [7]. Several studies have focused on predictive analysis of bike-sharing data [8], [9], [10]. Few of them tried to forecast the demand for bike usage using Machine Learning models [11], [12], [13].

Even though many types of researches were conducted regarding bike-sharing systems, very little is known about dockless bike-sharing systems and private bike usages. Moreover, when analyzing bike user behavior many research papers concentrate on only one variable, while in our case, we have analyzed user behavior considering the weather, social events, public holidays, pollution, seasonality, and city landscape. We have conducted a temporal and spatial analysis to understand bike usage patterns. In the second part of the thesis, we have used deep learning and machine learning approaches to forecast number rides in a given time interval. We have received the best results with long-short term memory neural networks(LSTM NN) model.

With the recent technological developments, a massive volume of data is collected by companies that provide bike-sharing services, hence there are more possibilities to use data when deciding on the location of bike lanes and for the prediction of the utilization of those lanes. In this thesis, a data set of bike mobility has been analyzed, general machine learning and deep learning methods have been used to predict bike demand at a given time interval, to provide the best places for creating bike lanes. This way bike lanes will be highly utilized by cyclists and will decrease traffic problems. Additionally, the master thesis will focus on the dockless bike user behavior and proposes an analytical framework to understand the behavior of bike users. We will analyze 6-month transportation mobility data of the Italian city Bologna. The dataset contains vehicles of different types such as

cars, buses, and bicycles. In this thesis, we have focused on the mobility of bicycles.

The dataset that is used in the thesis is the property of SRM - Reti e Mobilità Srl[1]. It was collected the following way. Each person downloads the BetterPoints application[2]. When opening the application user chooses whether they move by bus, by car, by bike or just walking. After that application starts sending data related to the users' position to the database. Location tracking is done via GPS (Global Positioning System) by the users' smartphone. The application takes advantage of both the American and Russian satellites or those of the GLONASS (Global Navigation Satellite System)[3] system which increases the accuracy of the user's position.

From the data on mobility, it is possible to reconstruct the map of Bologna, calculate the paths that users take, while still maintaining the anonymity of the user. Once the map of Bologna has been reconstructed, subnets can be calculated. Another important thing to calculate is the density on the streets of Bologna. The density is assigned knowing how many times a road is crossed. Moreover, the most attractive points of the city for bike rides will be analyzed and reported. We will look at how people spread out from these locations as well and which streets are the most utilized ones. This information will be in great help for the bike-sharing companies to find the bikes which are not being used due to some reasons.

Next, a temporal analysis is done by using different plots. Different quantities will be analyzed, such as the average speed of each trip, the lengths of the journeys, and travel times. The units of measurement of these three quantities are [minutes], [m/s], and [m] respectively for time, speed and space. With the statistical analysis, we try to see how the mobility changes in the various months; in particular I will focus my attention on May and August, since during May universities and schools are open and many people will use bikes, while in August all the schools will be closed and many people will go on a vacation to other countries, hence we will see a large a change in the mobility of bicycles.

Finally, in the predictive analysis part, we experiment with different machine learning models and try to get accurate predictions using historical data of bike usage and various features that affect bike usage such as air temperature, precipitation, and public holidays. The model was trained for 30-minute and 60-minute intervals, additionally, K fold cross-validation was used to improve the results.

The structure of the thesis is organized as follows.

1. In Chapter 2 works of similar topics are presented, including data analysis of shared

[1]SRM is the local Authority for Public Transport in the Bologna area, http://www.srmbologna.it/

[2]BetterPoints Ltd, https://www.betterpoints.ltd/

[3]Information and Analysis Center for Positioning, Navigation, and Timing, https://www.glonass-iac.ru/en/

bicycles in a city, traffic prediction in bike-sharing programs, and prediction of the number of trips done by bikes with Machine Learning algorithms.

2. In Chapter 3 the data that is used in this thesis is described and data transformations are outlined.

3. The Descriptive analysis of the data is presented in Chapter 4, where we conducted temporal and spatial analysis and analyzed the effect of weather and different events on the usage of bikes.

4. Prediction models, metrics, and results are presented in Chapter 5.

5. In the last chapter of the thesis, our conclusions and ideas for future works are mentioned.

2 Literature Review

In this chapter of the thesis, we examine academic works that focus on bike analysis patterns in specific cities. Firstly we bring out papers that have investigated the behaviors of the users in bike-sharing systems. And secondly, we looked at papers where authors tried to predict with different models how to distribute bikes in the city for easier usage, predicting pulse and traffic flow.

2.1 Analyzing Bike-sharing systems

With the increasing problems over global warming and dangers with hyper-urbanization in many cities with a high population, a lot of actions are conducted to make people use a more eco-friendly way of transportation such as bikes. Bike-sharing seems to be a possible solution for the highlighted problems as it does not harm nature, reduces traffic congestion and pollution. Currently, dockless bike-sharing programs are becoming more and more popular all over the world. As this type of bike-sharing is not restricted by station infrastructure, people who are using dockless bike-sharing can park the bike anywhere they desire. On the other hand it is harder to forecast where exactly the bikes will end up and how it will be used in the future.

Many researchers are analyzing patterns of bicycle usage mobility in order to understand how bike usage can be improved. Froehlich et al [4] investigated behaviors of bike users across different locations, neighborhood and times of the day. In the research, paper authors analyzed 13 weeks of bike usage from a shared bike system called Bicing, in the city of Barcelona. There are three main contributions of the paper: it demonstrates the potential of using data gathered from the bike-sharing activities in order to analyze and get information about the dynamics of the city, explores the relationship between the usage of bikes and city behavior and geography and finally studies usage patters of bike stations. While in the following study about the preferences of bike users Liu et al [5] concentrated more on whether the road network characteristics influence the choice of the path. The research allowed them to determine which road attributes influence bike users' path choices. For example, a cyclist is using a safe and motorized road more even if that means they will spend more time on the trips. Additionally, the study shows that bike users try to not use intersections, possibly the reason is that users try to decrease waiting times at the intersections.

Another research concentrated on bike-sharing operations in the city of Cork, Ireland [14]. One of the main motivations to study bike usage in Cork for authors was the desire to understand how these bike-sharing systems are performing in relatively small or medium

European cities. In small cities, there is a clear visible pattern of trip duration, which are on average very short. The research was also conducted in regard to weather conditions and how it affects the usage of bikes in Cork, showing that trips are shorter during rains and longer trips are more done during sunny days. At the end of the paper, the authors try to connect the results with the analysis done in bigger cities. Similar analysis was done by Handy at el [15]. In the paper author's analyse factors that are correlated with the bicycle commuting in six small cities in United States of America. The mentioned research papers are valuable for our study as the city of Bologna can be considered a rather Medium city with having a population of around 300K.

Numerous researches were done in order to understand the effect of weather on the use of bicycles. Nosal and Miranda-Morena [1] studied how weather affects the usage of bicycle facilities in North American cities. Authors found that precipitation in a single hour might significantly affect number of bike rides. They noted also that cycling during weekends are more affected by weather than during weekdays. El-Assi et al [2] studied how weather affects bike sharing demand in the second biggest city of Canada, Toronto. The findings imply that there is a significant correlation between air temperature and bike usage. Authors also investigated how built environment affects bike demand, concluding that bike infrastructure plays a major role in increasing popularity of bike usage. Nankervis [3] studied weather's long-term and short-term effect on bike trips. Study tries to tackle the notion that short-term weather or daily temperature and long-term weather or seasonal conditions affect bike usage significantly. Author concluded that both assumptions were upheld by the study, but the effect is weaker than it is generally assumed. Finally, Sears et al [16] discuss how seasonal factors affect with bicycle commuting. The results of the study confirm that there is a high correlation between weather and bike usage.

In Melbourne and Brisbane, researchers tried to quantify the factors influencing bike-share membership by preparing a questionnaire and analyzed the results by using logistic regression found out that various factors impacting on bike-share membership in Australia [17]. The results of the paper show that the distance to the closest docking station is highly correlated with the membership and the authors indicated that the result confirms another research which was conducted a bit earlier [18], [19]. Another important point that was mentioned in the paper is that 30% of the non-members are concerned with the safety of riding bikes. As people who are not using the membership cards are more concerned about the situations on the roads of the city and do not think it is very safe for riding a bike regularly. Paper also shows that people who are members of the system have a significantly higher income than other groups.

Several studies investigated how bicycle usage impacts air pollution level [20], [21],

[22]. In general, studies imply that using bicycles instead of cars and buses will reduce air pollution exposure. On the other hand Strauss et al [23] highlight that the air pollution level is highest near the most utilized cycling facilities. In this master thesis we will analyse how different air pollution indicators change during 6 months period of our observation, we will also analyse whether there is a short-term effect of bike usage on air pollution level.

Other authors [24] explored the bike-sharing travel time and trip by gender and day of the week. The result of the study suggests that demand for bike usage is generated in the residential districts, while the biggest hubs are train stations. The paper shows that there is a big difference in the sense of distance and trip duration between men and women. Additionally, the paper shows that women are using bikes during weekdays more, while men use it on average more during weekends. Finally, another paper that examines the behavior of bike usage in a city focuses on how social behavior will help in planning and designing policies in transportation [25]. The authors study a shared bicycle system called Velo'v which is located in Lyon, France. The paper tries to reveal usage patterns of bikes, with regards to social studies of transportation, and for that author uses signal processing and data analysis.

In this part, we have analyzed different research papers, the main focus of which was the behavioral analysis of users. Multiple authors are analyzing and tackling various problems regarding bike usage, but are focusing on just one or two main points. In this master thesis, we will try to combine multiple features that are affecting bike users' behavior. Also, we will pay more attention to the spatial and temporal dynamics of bicycle usage in order to provide a holistic picture of the usage patterns of shared bikes.

2.2 Predictions using bike data

A lot of researches has been conducted about predicting different scenarios by using bike data. Most of the authors were trying to use different Machine Learning algorithms and create different models to predict the usage of bikes and make them more comfortable for users. Vogel et al [26] analyze operational data from bike-sharing systems to understand activity patterns and to use these patterns in order to observe the distribution of bikes. Later, a data mining process is introduced to plan and manage the imbalanced distribution of bikes.

Li et al [8] predict a number of bikes that are rented from each station and try to understand how the relocation can be done in advance so in a bike station where the demand is higher will not run out of bikes. The paper proposes a hierarchical prediction model in order to predict how many bikes will be rented from or returned to each station so that the workers can relocate the bikes in advance. One of the main problems that the

authors did not tackle in the study is different unusual factors when predicting traffic.

In contrast, Xu et al [26] took into account different factors when predicting bicycle traffic flow. The study uses a hybrid prediction model which is combining clustering with support vector machine instead of a single prediction model. This gives more accurate prediction results. Additionally, the proposed solution is now being used in the city of Hangzhou, China. The same problems were also tackled by Zhang et al [9]. The authors tried to predict the destination and arrival time of each individual bicycle trip which can be later effectively help the companies to move bikes on time and to the correct (under-supplied) station. Based on the analysis study provided two new regression models which successfully predict trip duration and station.

The following research done by Duc-Nghiem et al [10] developed a model for predicting the facility choice of cyclists between on-street facilities and off-street facilities based on bike users' behavior. As in some cases, cyclist rent avoiding using new bike facilities, like bicycle lanes. The authors tried to understand and analyze what are the reasons for not using the new bike facilities. The initial analysis showed that the bus stop existence, the width of the sidewalk and the type of bike are the main variables for the prediction model. A framework for predicting bike lane usage was presented and it is suggested to have a good predicting ability. While research done by Singhvi et al [27] used CitiBikes websites data were they predicted pairwise demand for New York city bikes. During the study taxi usage in addition to weather data was used to predict activity trips. The analysis was done only for the rush hours of 7:00 AM – 11:00 AM. Authors provide a model that should assist bike-sharing in the "macroscopic level".

Recently, a number of studies have investigated dockless bike-sharing systems in order to understand the usage of such systems. Some papers discuss the difference between docked and dockless bike sharing systems analyzing how the bike sharing pattern differ between the two systems [7], [28], [29]. Other authors try to analyse spatial patterns of such systems, as without stations bicycles can be left anywhere in the city, which also raises the question of redistribution of bikes during the week [6], [30], [31].

Several attempts have been made in order to forecast bike demand using deep learning approach. Some of them experimented with a large-scale datasets and tried to predict demand for different time intervals using only historical bike usage data [12], [13]. Zhang et al [11] tried to use Long short-term memory (LSTM) model to predict number of rides considering public transport usage as well. Authors' experiments provide an accurate predictions results in comparison with the baseline. In our knowledge no study was conducted predicting number of rides considering not only historical bike usage data but also air temperature, precipitation, public holidays and special events such as protests and strikes. This paper tries to fill in the mentioned gap.

This part combines research papers that were focused on predicting different instances with the bike movement datasets. Most of the papers are concentrating on redistribution of the bikes in different locations, and trying to predict in which place there will be a surplus of bikes in which place a shortage of bikes. Some of the papers are analyzing more specific tasks such as whether on-street or off-street facilities will be used by cyclists. There seems to be a lack of researches on predicting dockless bike location, which is more complicated due to the fact that bikes can be left anywhere, and there is no finite number of places were the bikes might be located. We focused more on that matter and created a model that will show on a map on a specific time where the bike will be located.

In this chapter, we have gathered related work that was already done in regards to user behavior analysis and predictions by using data collected from bike-sharing companies. Though there is a great number of research papers regarding bike-sharing with dock stations, there was little research done regarding the analysis of user behavior regarding dockless bike-sharing services. In most of the cases, authors are analyzing one or two features that are affecting bike usage, while in this master thesis we have tried to combine more features that we think affect the behaviour of bike users. Additionally, in this thesis will forecast bike demand using a deep learning approach. We believe this will help to tackle the problems with traffic and with the uncertainty whether the new bike lanes will be utilized or not.

3 Data Description

This chapter of the thesis gives information about the data set we used for the analysis. Mainly we describe from where the data set was collected, how it was prepared and cleaned, and which features were used during the analysis and predictions. Finally, we have performed a descriptive analysis to describe all the features of the data set.

3.1 Datasets

3.1.1 Bella Mossa program.

The main dataset used in the master thesis contains mobility data of different transportation means for the period from April 1, 2017, till September 30, 2017. Bella Mossa is a program that promotes a healthy lifestyle and sustainable mobility. It gives the user a chance to win various gifts and discounts as a reward for using a more sustainable and healthy means of transportation. Participation in the program is very easy, a user just needs to download the application and start it whenever they go out for a walk, use bikes, trains, buses, or even when using carpooling. After that application sends data related to the users' position to the database. Location tracking is done via GPS by the users' smartphone. During 6 months of the experiment, there were over 15000 unique users of the program and 3.7 million km was covered by them.[4] For security reasons users' personal information is not saved and is not available in the data set, thus we will not be able to distinguish and analyze different patterns considering users gender or age. Worth noting that each trip is separated by a unique identification number (ID), which does not allow users to be identified on different days. Figure 1 shows the study area of the thesis.

The data includes information about car usage, bike and bus usages, additionally there is an information about trains and also about the mobility of pedestrians. In total there is a data of 15000 active users. The Table 1 shows all the variables from the main data set that we have received.

In the thesis, we focus on the behavior of the bike-users and analyze how they behave during different periods, during different months or seasons, during different weather conditions and different events. We are also interested in forecasting the number of rides during a certain period to understand how utilized the bikes will be and how busy specific streets will be, hence we will alter the data in many different ways to answer these questions. We will group the data points based on unique Activity IDs to analyze factors affecting the number of rides at a certain time and will use raw data with all the points to reveal which streets and parts of the cities are mostly utilized by bike users.

[4]Bella Mossa official website, https://www.bellamossa.it/

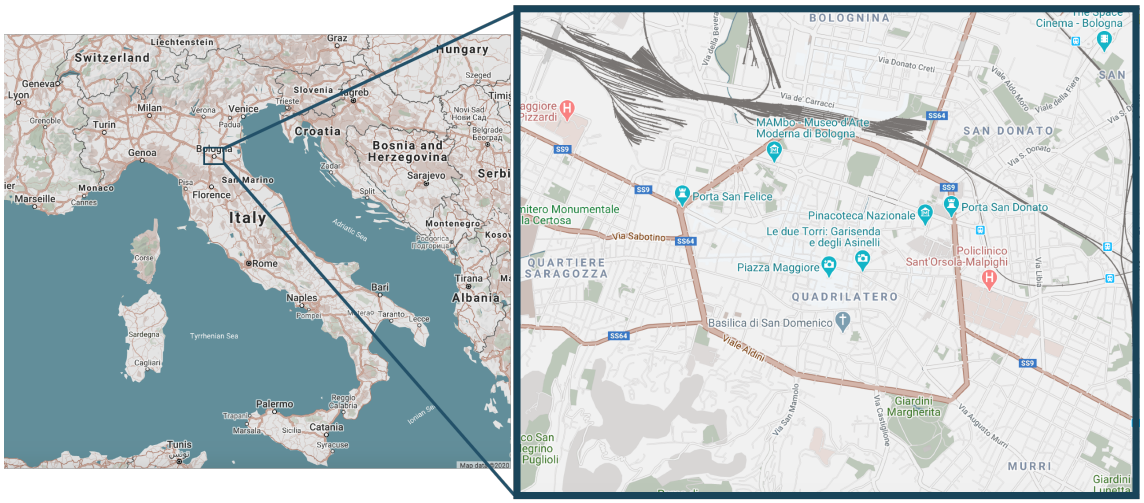


Figure 1: Study area in the city of Bologna

3.1.2 Alternative datasources

Besides, to the main dataset, we have used several alternative datasets for analyzing bike usage and predicting the number of bike trips.

First, have downloaded historical data about weather, precipitation, and wind for the observation period for the city of Bologna from Dext3r website. Dext3r is an application which collects the weather data which recorded and managed by Regional Agency for Prevention, Environment, and Energy (ARPAE) of Emilia-Romagna, Italy.[5] Weather data contains information about the daily and hourly average air temperature above 2 meters from the ground. Wind data contains information about the average wind speed above 10 meters from the ground. Data is collected daily and hourly. Precipitation data shows cumulative precipitation over 1 hour period and a day.

Secondly, we have used publicly available information about the holidays in Italy and public events in the city of Bologna[6]. The holiday data set contains information about all public holidays, national celebration days, civil solemnities, and celebrations in the city of Bologna. Thirdly, we have used the information about pollution from the website of the Ministry of Economic Development in Italy[7].

[5]Regional Agency for Prevention, Environment and Energy of Emilia-Romagna [website], https://simc.arpae.it/dext3r/, (accessed 13 January 2020)

[6]The Council of Ministers [website], http://presidenza.governo.it, (accessed 28 January 2020)

[7]Ministry of Economic Development in Italy [website], https://www.sviluppoeconomico.gov.it/index.php/it/, (accessed 17 February 2020)

Variable	Description of the Variable
Activity Id	This parameter identifies the individual trip
Activity Type	Declaration by the user with what moves
Time	Identifies the date in year, month and day format with the time next to it.
Latitude	Latitude
Longitude	Longitude
Accuracy	GPS accuracy.
Speed	Speed of the person.
Identified Type	The program identifies established with which vehicle the person moved. Sometimes it happens that it does not recognize which type was used for example when someone used the bicycle but then halfway through took a bus.
Identified Confidence	This value tells how confident we are that there is a match between the statement that a person moves and the statement calculated by the program.

Table 1: Description of all variables of the main data set

Four main air pollution indicators were analysed during the period of the experiment:

- Particulate matter (PM)
- Ozone (O3)
- Nitrogen dioxide (NO2)
- Sulfur dioxide (SO2)

Based on the 2005 World Health Organization(WHO) Air quality guidelines[8] these four indicators are evaluated to understand the air pollution level globally.

Finally, we have gathered information about strikes and protests in Italy. Dataset was downloaded from the official website of The Ministry of Infrastructure and Transport in the Republic of Italy [9]

[8]World Health Organization [website], https://www.who.int/, (accessed 15 February 2020)

[9]The Ministry of Infrastructure and Transport in the Republic of Italy [website], http://scioperi.mit.gov.it/mit2/public/scioperi/ricerca

3.2 Data Cleaning and Preparation

For our analysis, we have used bike related data only and removed information about all other transportation means. We have, as well, removed features like accuracy, Identified Type, and Identified Confidence as those were not providing any values during our analysis and where contained many missing values. In total, the new data set containing information about bicycles had 72,398,780 data points and 320,118 unique trips. Speed column had over ten thousand missing points, as it was less than 1% of the original data it would not have affected our analysis and it was decided to fill these missing points with values from the observation that proceeds before them. Furthermore, we have added variables from the "Time" variable from the original data set, the new features are "Date" - date of the trip, "Hour" - the hour of the trip, "Weekday" - shows the day of the week of the trip.

To calculate the number of rides we have counted the number of unique Activity IDs and as can be seen from Table 2 the most number of rides were done in May and the least number of rides were done in August. The most popular day in sense of riding a bike is Wednesday and the least popular is Sunday. A more detailed analysis is provided in Section 4.6. Also from the mentioned table, we can see that the average speed for trips was around 13.5 km/h and the average trip distance was around 7 minutes.

As in the dataset, we had information about Latitude and Longitude of each point at a certain time, it was quite easy to analyze the directions of trips, how exactly they have spread in the city and which streets were the most utilized ones. Furthermore, it provided a possibility to create a density plot for each hour during the observation period and analyze how streets are utilized during different times of the day.

Characteristics	Value
Number of unique trips	320,118
Number of unique points	72,398,780
Observation period	April 2017 - October 2017
Average speed	13.5 km/h
Average trip distance	7 minute
Most popular month	May
Least popular month	August
Most popular day of a week	Wednesday
Least popular day of a week	Sunday

Table 2: Descriptive statistics from our data set of Bella Mossa program

For the predictive analysis, the categorical features had to be altered, so that it will be easier to use them in the machine learning models. For this, dummy values were created from the available features. More information about predictive analysis is provided in Section 5.5

4 Descriptive Analysis

In this section, we will present results obtained from the analysis of the Bella Mossa data. Primarily, we will look into daily, monthly and seasonal trends of bike usage, secondly we will do spatial analysis of the data to understand in which part of the city most of the bike trips are happening and how bike users are spreading out from those places. Next, we will analyse how different factors, such as change in air temperature, precipitation and public holidays are affecting the usage of bikes. Furthermore, we will also take a look at the pollution data taken from 2 different domains and will check whether the pollution level is changed due to heavily usage of bikes instead of other transportation means. Finally, we will analyse whether events such strikes and protest in transportation services are affecting the number of bike rides or not.

4.1 Temporal analysis of bike usage

We have aggregated data based on number of unique rides for the whole period of experiment. We start with the analysis of monthly usage of bikes, then will look closely into more detailed view and will look into weekly and daily trends. Our data is available for six months only, from April till the end of the September, therefore we can also analyse the change in user behaviour during Spring, Summer and start of Autumn.

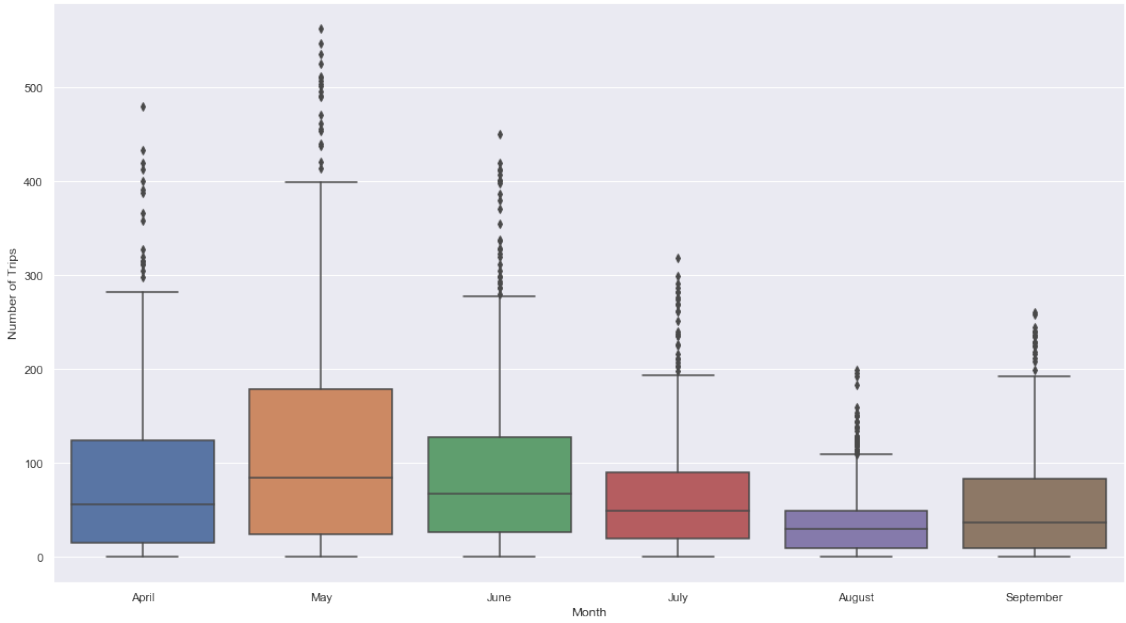


Figure 2: Number of Trips Aggregated by Month. Source: Author's calculation from Bella Mossa data.

Total number of bike rides during 6 months period was 319,926, out of which 25% is

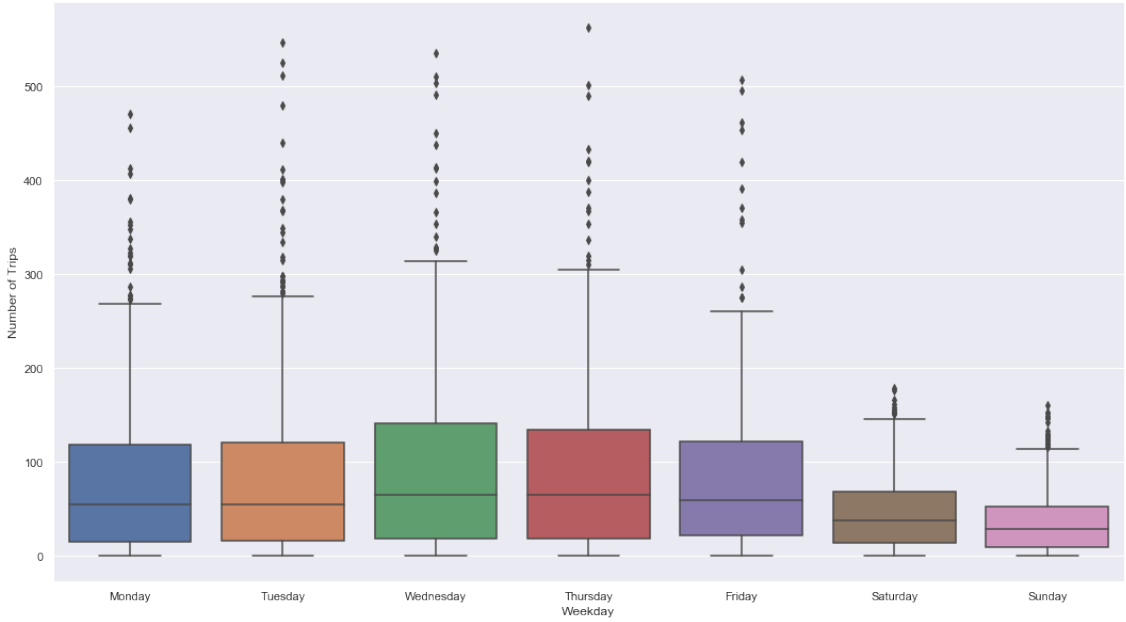


Figure 3: Number of Trips Aggregated by Weekdays. Source: Author's calculation from Bella Mossa data.

done in May, which is the most popular month for riding bike based on our data. From the Figure 2 we can see that the number of rides reaches its peak during May and then it gradually drops down, reaching its lowest level during August. There was a around 40% drop in rides from July to August. Additionally, in August only 30% of May's rides were done. Figure 2 also shows that number of rides is decreasing through Summer, this can be explained as more and more people are leaving the city and going for a vacation or students going back to their home places during summer. At the start of September number of rides starts to increase, it was 28% higher than number of rides done in August, but it never reached the level of rides done in Spring or even in mid of Summer. One way this can be explained is that people started to get bored with the application and the experiment and were no longer keen to use it to track their rides, or they started to used some other apps which lead to decrease in rides during September. Start of Autumn is considered one of the most popular seasons for biking, especially in cities where most of the habitants are students and use bikes a lot for traveling from home to university buildings.

Next we looked into rides aggregated by weekdays. From the Figure 3 we can conclude that bike rides are mainly taking place during weekdays. During weekdays 84% of total rides are done, and only 16% is done on weekend. This probably is connected with bike users going to work or to studies and coming back home. The most popular day for bike usage in average is Wednesday and the least popular in average is Sunday. Overall on

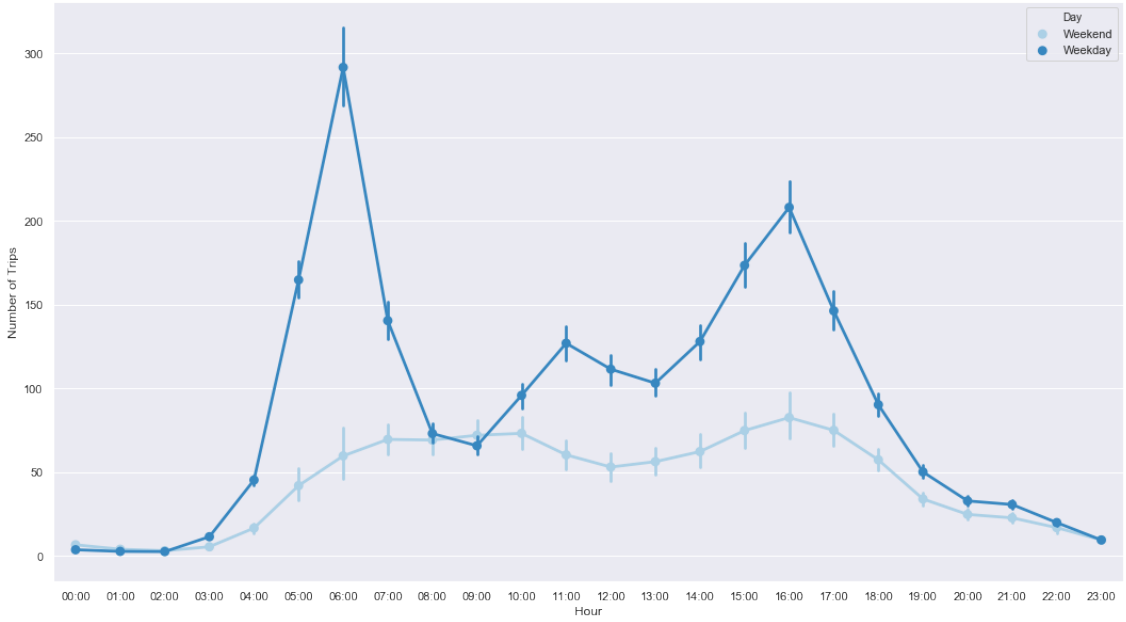


Figure 4: Comparison Between Weekends and Weekdays. Source: Author's calculation from Bella Mossa data.

weekends bike usage drops heavily implying that people stay at home or use other means of transportation.

The assumption of using bikes mainly for going to studies or work and coming back home is also confirmed with the daily split of the data.

From the Figure 4 we can clearly see that the most of the rides during the day are happening from 06:00 - 07:00 and from 16:00 to 17:00 this is mostly connected with people going to work in the morning and coming back from work to home.

As we have seen from Figure 3 weekends are quite unpopular in terms of bike usage and similarly the usage during a day should be different on weekends and on weekdays. In Figure 4 we have plotted how aggregated hourly data separately for weekends and weekdays. This shows a clearer picture on when do people usually use bikes during weekends and when they use it during weekdays.

A bit more detailed information about the split of the trips during different weekdays and different time of a day is presented in the Figure 5. We have aggregated Number of Rides by weekdays and hour of the day and summed the values.

From the Figure 5 we can conclude that most of the rides are done from 6:00 to 7:00 during weekdays, with the largest share of rides being done on Thursday. The smallest number of rides are happening in midnight. It also worth to note that number of rides during midnight increases on Friday, Saturday and Sunday, implying that many people spend more time outside enjoying their weekends.

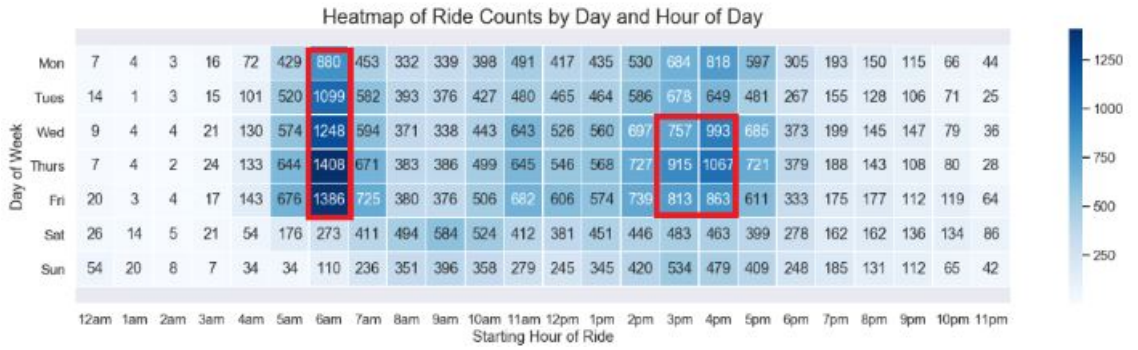


Figure 5: Heat map of Rides During Weekdays. Source: Author's calculation from Bella Mossa data.

4.2 Attractive points for cycling

In this subsection we will ask what are the main city hubs for bike users and in which parts of the city people are mostly using bikes. To answer this we will begin plotting all the points from our data on the map of Bologna. We will try to create a heat map and show which places are the most popular ones for the bike users in the city of Bologna.

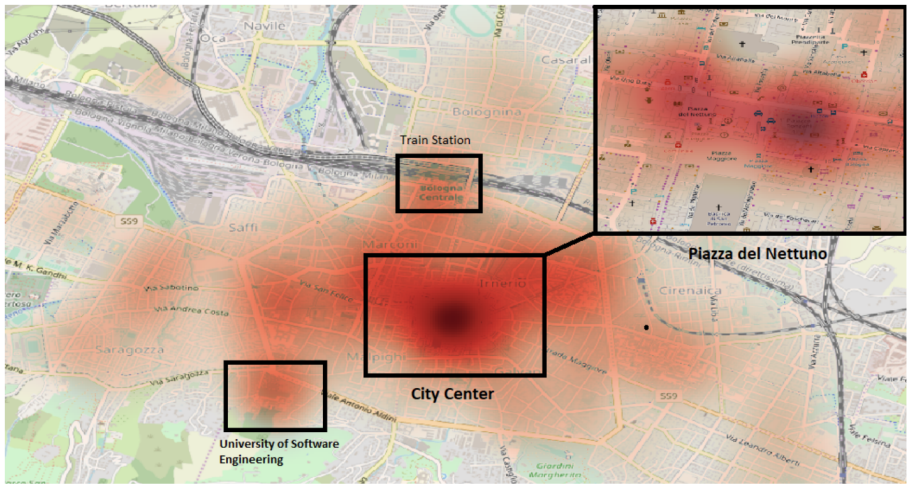
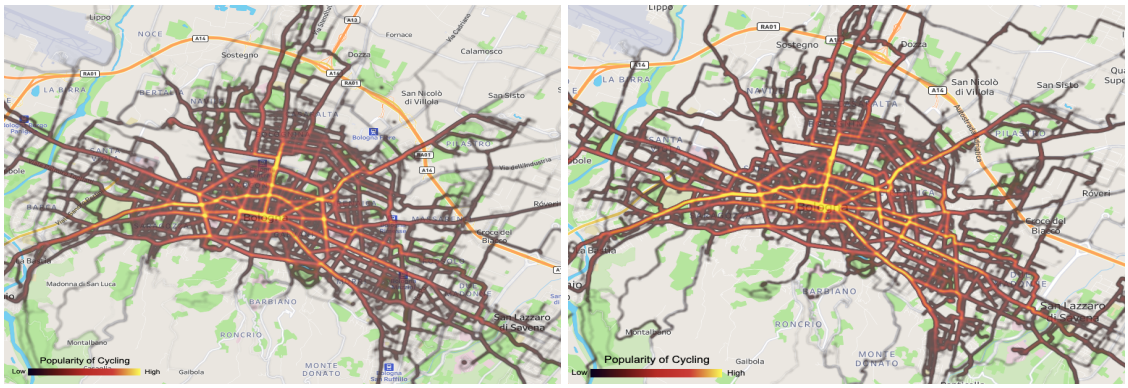


Figure 6: Density Plot of Bike Rides. Source: Author's calculation from Bella Mossa data

70% of total rides are done in 3 main places and we can also see in Figure 6that the most attractive points are City Center, with the main point being streets around Piazza del Nettuno. Second attractive point is the University of Bologna, Department of Software Engineering, as the city is considered as a student city it makes sense that this is one of the most attractive hub in the city. Third most attractive point in sense of bike trips, similar to the findings of Zhao et al [24], is the train station area. Many people commute for work

or studies and they use bikes to arrive to train station or to ride from train station to their workplace or to university.



(a) Density of bike rides. May 2017 (b) Density of bike rides. August 2017

Figure 7: Difference between trips done from Train Station during May and August. Source: Author's calculations

Next we analysed which streets are the most used by bike users. For that reason we have created a density plot out of all the points of the main data set. The three most busiest streets are the following ones Via Rizzoli, Via dell'Indipendenza and Via Sabotino. Furthermore, we compared usage of bikes during May and August, as these were the most popular and least popular months respectively.

(a) Density of bike rides. City Center, May 2017 (b) Density of bike rides. City Center, August 2017

Figure 8: Difference in density of trips done during May and August. Source: Author's calculations

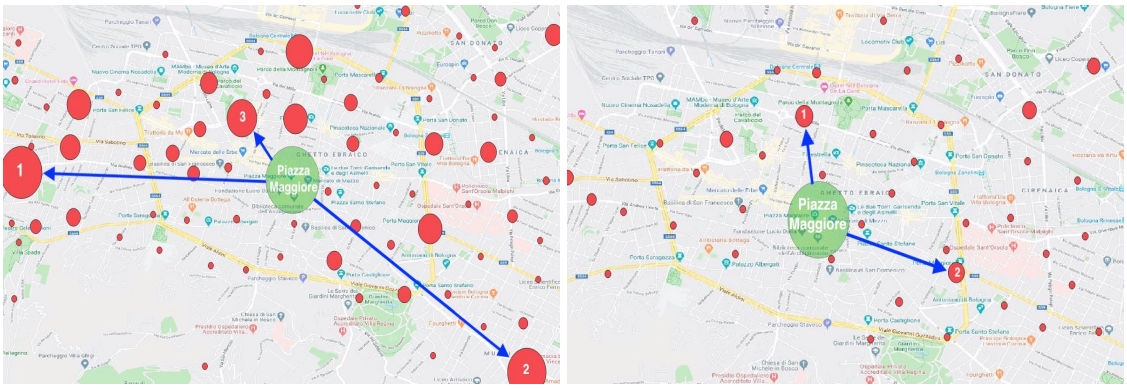
In both Figure 7a and Figure 7b we can see that density of the streets are similar. Mostly streets in city center are heavily used. This means that in case the city officials would like to start creating new bike lanes, they should certainly start from the areas near the above mentioned attractive hubs and three bustiest streets. This way city officials can be sure that the bike lanes will be utilized.

When zooming in more into the city center we again notice similar pattern in street usage with Via Rizzoli and Via dell'Indipendenza being the most heavily used ones. We can also notice that during August streets surrounding Bologna university buildings are used lightly, which is also logical as the Universities are closed during August.

4.3 How people spread out in different directions from the attractive hubs?

In this section we try to understand how bike riders spread out from the attractive hubs mentioned on the Figure 6. The locations are aggregated by the number of trips started from the attractive point and ended in the final destination. The more trips ended in the same location the larger is the circle in the below presented figures.

First, we will start analysing how users spread out from the Bologna city center. Figure 9 compares how users spread out from City center during May as shown in Figure 9a and during August in Figure 9b. It is clear that the pattern is changing heavily, during May we can see many destinations which are outside the city center, indicating that people are going back to home after spending some time in the city center. Most of the trips are ending in residential areas in Costa Saragozza and Murri neighbourhoods. In contrast during August most of the rides are ending inside city center. We should also not that there are very few trips starting from city center in August, only 54, while in May number of trips starting from May almost reach to 1500.

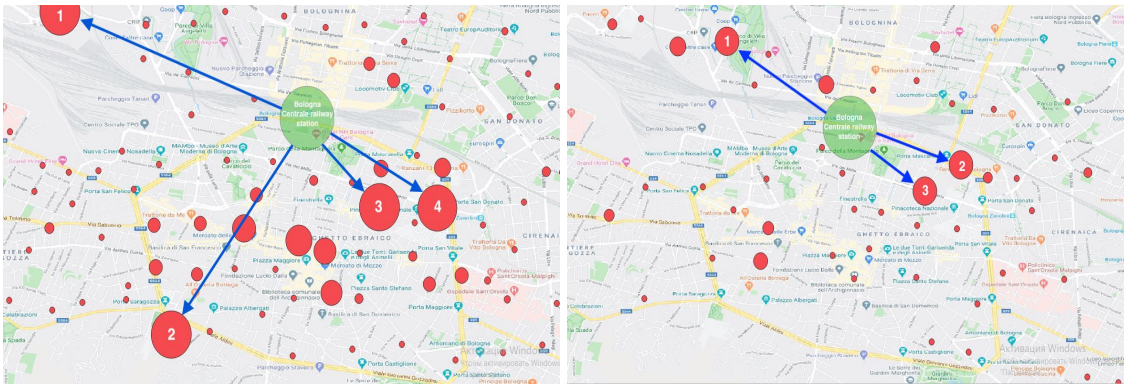


(a) May 2017 (b) August 2017

Figure 9: Difference between trips done from City Center, Piazza Maggiore during May and August. Source: Author's calculations

From train station bike users usually go by four main directions. The most popular direction is to the Ospedale Maggiore Carlo Alberto Pizzardi, which is the main hospital in the city of Bologna, and mainly people who commute from different places to work in

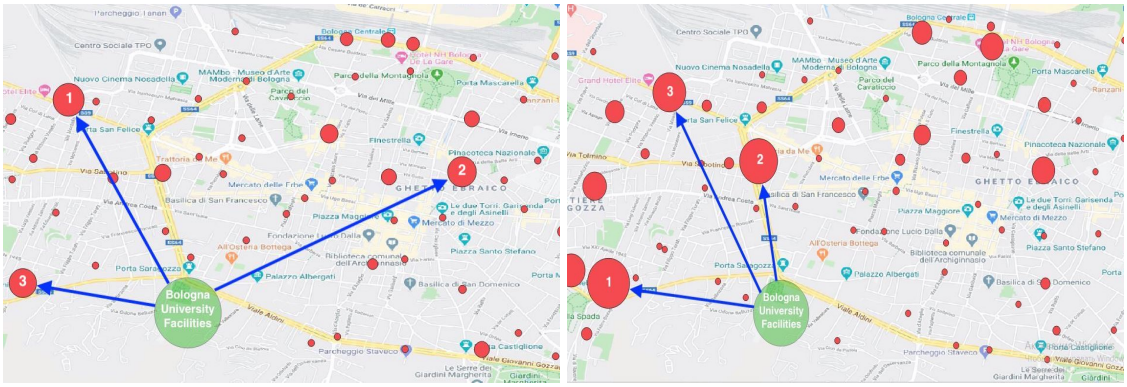
there. Next main direction is Bologna University School of Engineering and Architecture, which means mainly students and university workers are using bikes to travel from train station. Third is located around street Irnerio. And finally, forth most visited place from Bologna Centrale train station is Department of Mathematics - University of Bologna.



(a) Rides from Train Station. May 2017 (b) Rides from Train Station. August 2017

Figure 10: Difference between trips done from Train Station during May and August. Source: Author's calculations

Figure 10b shows how this patterns change during August. Besides the fact that there were much more smaller number of trips starting from train station (523 vs 2660 during May), we can clearly see that the final destinations change as well. There are just a few trips to he Ospedale Maggiore Carlo Alberto Pizzardi and to the Bologna University facilities. The most of the trips end in the Centro Lame a large shopping mall just a bit away from the city center. Other popular destinations from train station in August are near Porta San Donato.



(a) a (b) b

Figure 11: Difference between trips done from University of Bologna Facilities during April and May. Source: Author's calculations

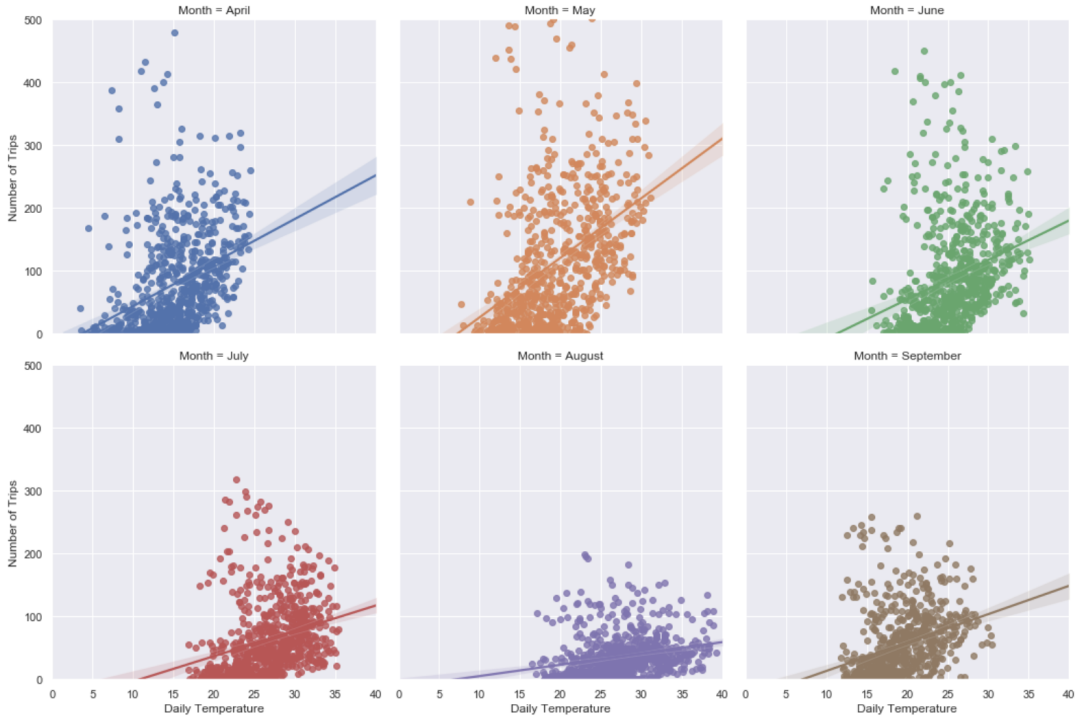


Figure 12: Average daily air temperature and the number of trips during 6 months. Source: Author's calculation from Bella Mossa data

Finally, we will look how bike users spread out of the third largest attractive hub from University of Bologna facilities located near Port Saragozza. Figure 11 compares how the pattern changes during months of April and May. During this analysis we looked into Spring months as number of trips starting from University facilities decrease heavily afterwards. Figure 11a shows that most of that start at the Bologna University facilities end either in city center either in residential areas of the city. This pattern does not change much during May, the only significant change is that more trips end near Porta San Felice.

4.4 Effect of weather and precipitation on bike usage

Usage of bike is significantly influenced with the weather conditions as mentioned in many researchers, thus we have also decided to look into the effect of weather on the usage of bikes [3] [2] [32]. In this subsection we will show the results of analysis of how air temperature, precipitation and wind speed affect number of rides done during the 6 months period we are observing.

As seen on Figure 12 weather plays significant role in the number of total rides done with bikes. During April, May and June it is clear that higher temperature makes users do more rides with bike, while when the air temperature is cold people prefer to stay at home or take another transport.

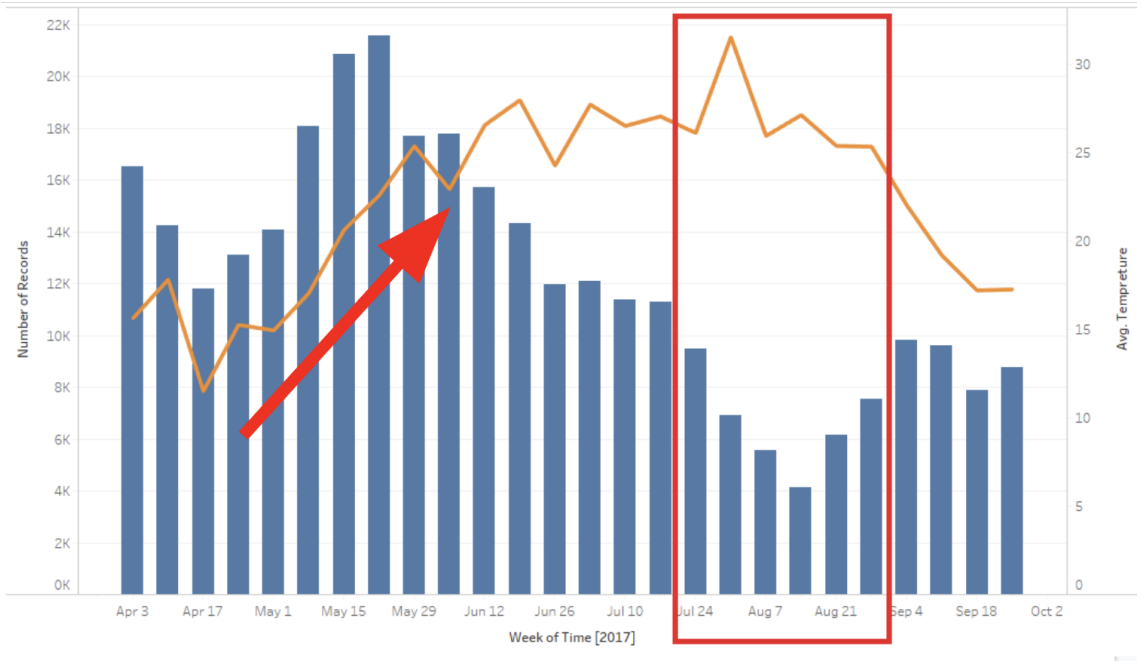


Figure 13: Average Daily Air Temperature and the Number of Trips. Source: Author's calculation from Bella Mossa data

On Figure 13 we see that from end of July till end of August the number of rides is dropping while the air temperature is staying at the same level, 26-27 Degrees. Another thing to note is that the highest number of rides during all 6 months are done when the air temperature was around 15-25 degrees. In summary we can confirm that bike rides are strongly dependant on air temperature which confirms findings of many authors.

When looking into cumulative precipitation data on Figure 14 it is clear that during the 6 months period of our observation most of the days were dry with no rain or snow. Only during September there were 7 days with some amount of precipitation while during

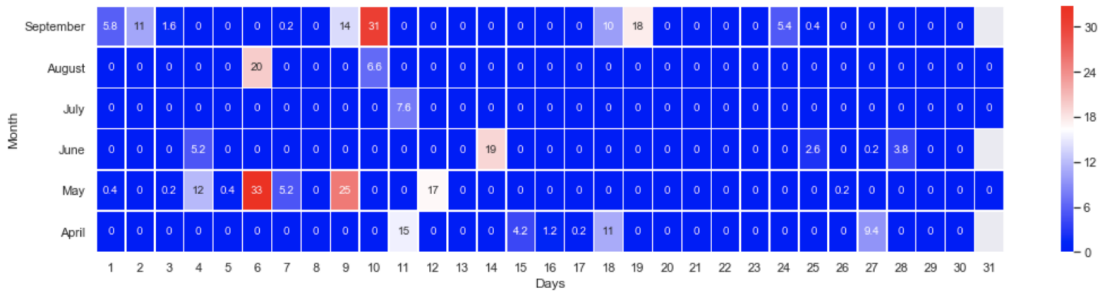
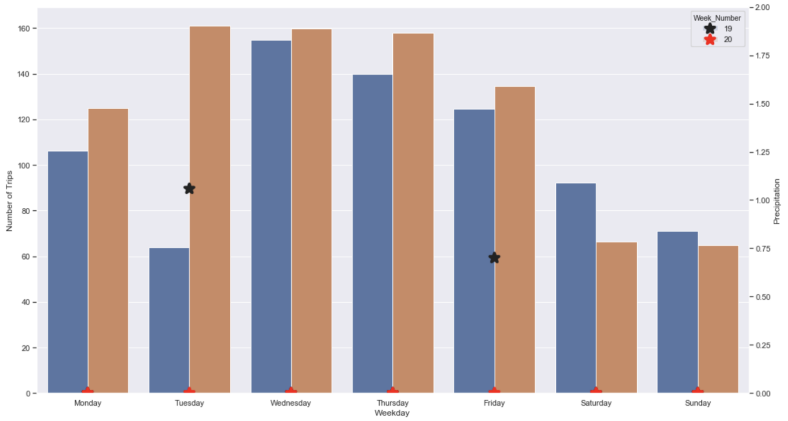
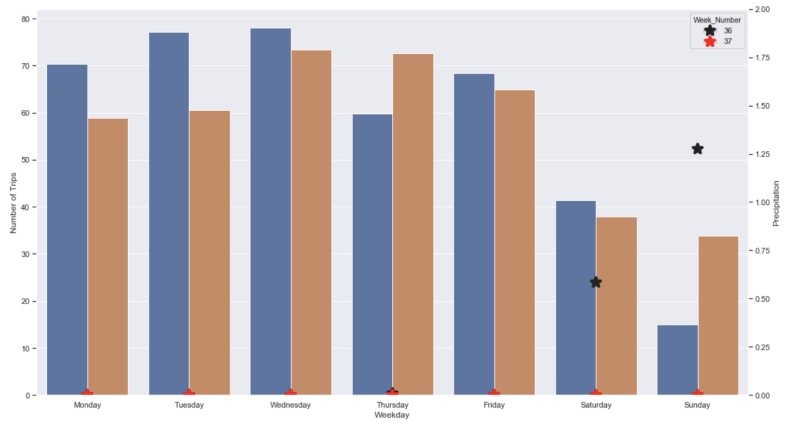


Figure 14: Cumulative Precipitation Data During the 6 Months Period. Source: Regional Agency for Prevention, Environment and Energy of Emilia-Romagna [website], https://simc.arpae.it/dext3r/



(a) Precipitation amount and number of rides during different weeks in May



(b) Precipitation amount and number of rides during different weeks in September

Figure 15: Precipitation amount and number of rides during different weeks in May and September. Source: Author's calculation from Bella Mossa data. Note: black stars on the plot indicate weeks during which there was some amount of precipitation.

other months number of days were significantly low. Thus, we decided to look into the weeks with the highest number of precipitation and compare it with the weeks before and after. This will let us know if the precipitation affects the number of bike rides or not. We decided to compare weeks during May (from May 08 - May 14 against May 15 - May 21) and during September (from September 04 - September 10 against September 11 - September 17).

Figure 15a shows the comparison done during the weeks mentioned about during the month of May. We can see that it was raining during Tuesday and Friday on Week 19 (May 08 - May 14) and there was no rain during next week. And this heavily affected the number of rides during the first week. On Tuesday number of rides is more than twice

lower than it is during next week on the same day. While there is a small difference between the number of rides during Friday we can see that it is not as significant as it was during Tuesday. When looking into hourly data we can see that on May 09 there were heavy rains from 2am till 9am and those rains affected number of rides largely, as have seen from Figure 4 most of the rides are done in the morning from 6am till 9 am. While on May 12 it was raining only for one hour around 3pm and this did not affect the number of total rides largely.

In Figure 15b we analyse how precipitation affects number of bikes trip during the above mentioned weeks in September. Similarly, we can see that during one day of the week on Sunday number of rides was significantly less due to rains, while on Saturday the number of rides was even higher. When looking more closely into data and checking at what time there was a rain we can noticed that on Saturday it was raining from 8pm to 10pm only and the drop in rides during these 2 hours was not significant. While, on Sunday it was raining from 2am to 11am, and morning rides were heavily affected leading to a twice lower number of rides during first week against next week.

Overall, we can confirm that during rains people are less likely to use bikes and more likely to use other transportation means such as bus or car.

Finally, we will analyse how the speed of wind affect usage of bikes. When plotting (Figure 16) number of trips and average wind speed we can clearly see that there is no correlation between two features, thus wind speed has not affected bike usage in Bologna during the observation time.

4.5 Effect of bike usage on pollution

Being one of the most ecological mean of transportation there is a notion that usage of bikes might decrease the pollution of air [21], [20]. We have analysed the change in three different indicators of air pollution during the start of the experiment and at the end of experiment. The data was gathered from two different sources. Overall, there was no positive or negative changes in the indicators implying that the usage of bikes did not have any impact on air pollution. We have also analysed if there was any changes in the indicators for separate streets which were being heavily used by bicycles. Similarly we could not find any correlation between air pollution and bicycle usage during the period of observation. We believe that the reason might be short amount of observation period and that the usage of bicycle will have a positive effect on air pollution only after long-term usage.

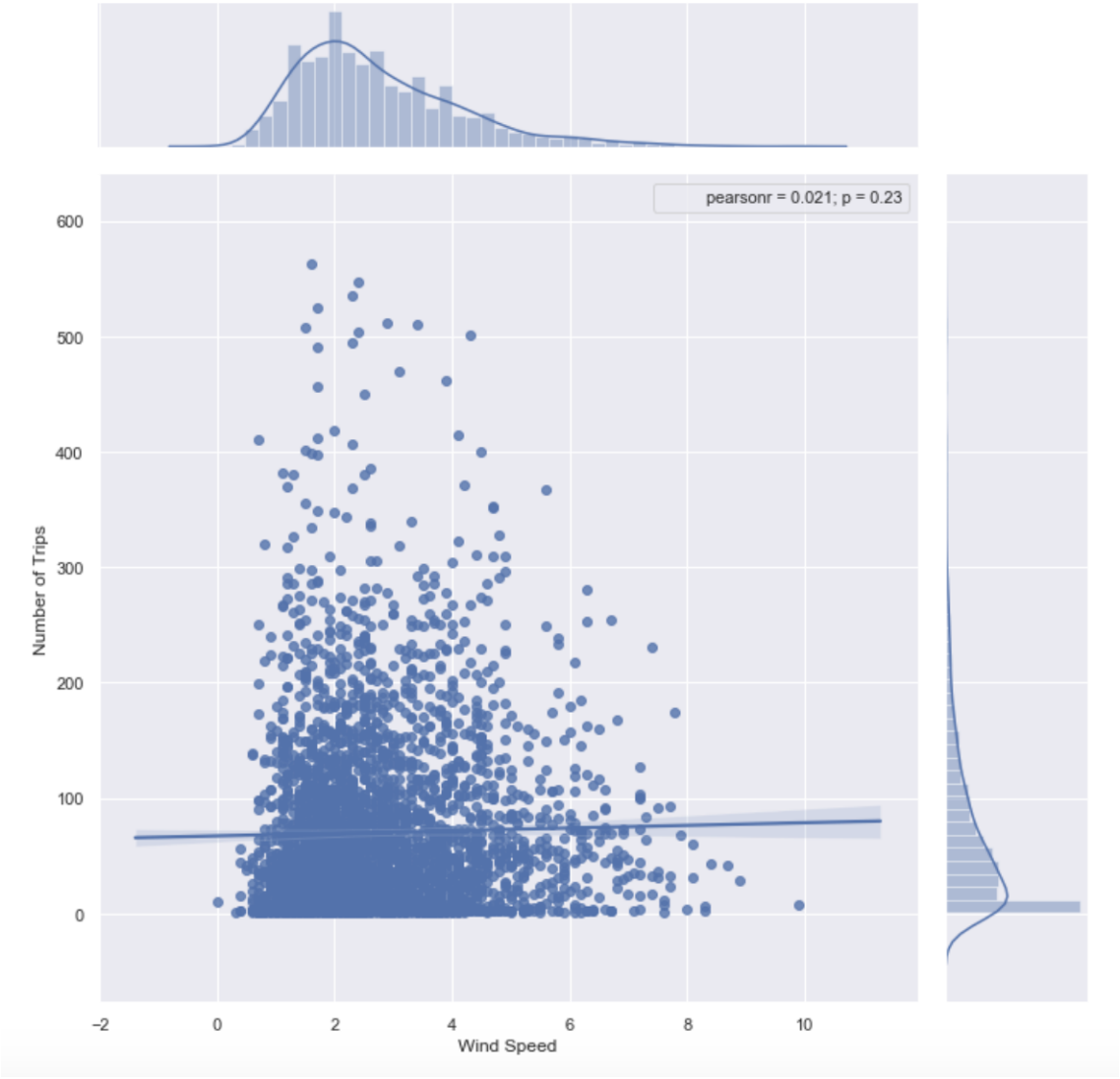


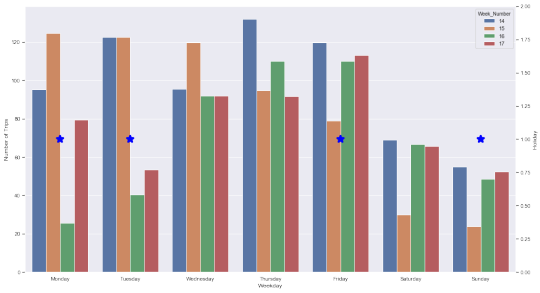
Figure 16: Average wind speed and the number of rides during the period of observation. Source: Author's calculations

4.6 Holidays and events

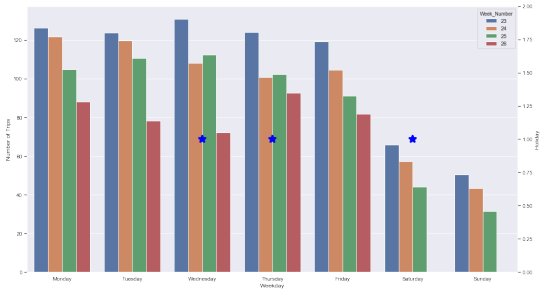
In the end we have analysed how different public holidays, special events such as protests and strikes, and national celebration affect behaviour of bike users.

During the period of the experiment there were 14 public holidays. We have analysed how behaviour of users change during these days and found out that significant difference was only during Easter holidays. On Figure 17a we see that there was a large drop during Easter holidays. Number of rides dropped by 35% on Good Friday, on Saturday and Sunday rides dropped over by 50% and almost by 80% on Monday and Tuesday. Clearly, Easter holidays are significantly affecting the number of bike usage.

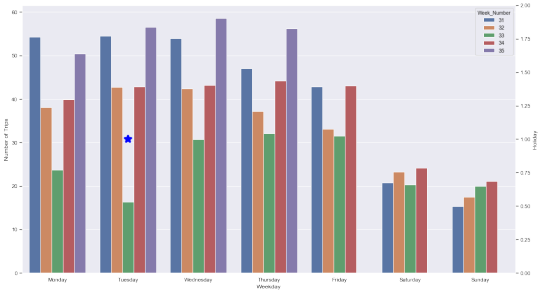
When analysing other holidays we see less significant changes. During June there



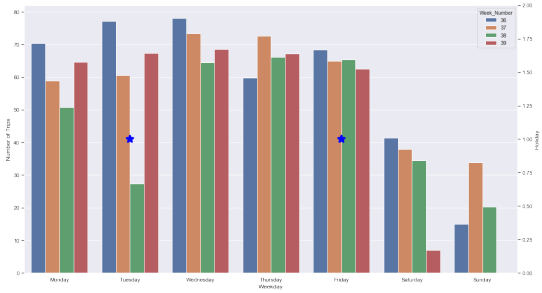
(a) Number of Trips and Holidays in April



(b) Number of Trips and Holidays in June



(c) Number of Trips and Holidays in August

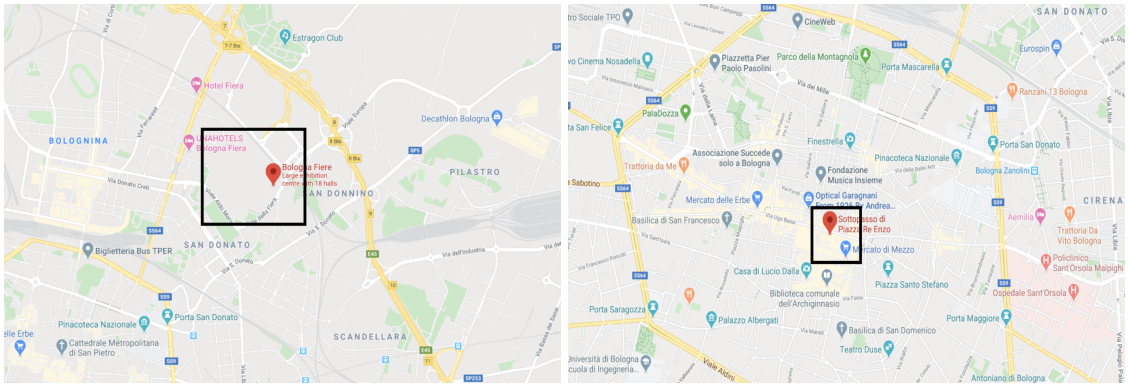


(d) Number of Trips and Holidays in Sept.

Figure 17: Number of Trips aggregated by Day of a Week for 4 months. Source: Author's calculations Note: Holidays are marked with a 'star'.

were 4 Holidays but none of them affected number of rides largely (see Figure 17b). On August 15 there were two holidays Assumption of Mary and Ferragosto this led to a decrease in number of rides by over 60% in comparison to the number of rides done one week ago and to the number of rides done one week after (see Figure 17c). On September 19th we see a drop in trips in comparison with the previous and next weeks, while on September 22 there is no change in the trips (see Figure 17d).

Finally, we have analysed how events such as protests, strikes and just large gatherings in the city as movie premieres affect the usage of bikes. The reason behind analyzing protests impact to the bike usage is in the following, during protests or strikes number of main roads are being blocked or closed, which can lead to increase in number of rides with bikes as more people will choose bikes instead of bus or car. On the other hand protests can also decrease the usage of bikes as people would either decide to stay home or take a walk, as the roads might be blocked for bikes as well. When analysing protest in Bologna during the 6 month period we have not noticed any significant change in the usage of the bike rides. It worth to mention that most of the protests and strikes were not large and were not lasting for a many days, it would end in 2-3 days the most. Additionally, we did not have an information whether the streets were closed or not and judging from the fact that number of rides did not change significantly we can conclude that during those 6



(a) Bologna Fiere (b) Piazza Re Enzo

Figure 18: Trips ending Bologna Fiere and Piazza Re Enzo were analysed

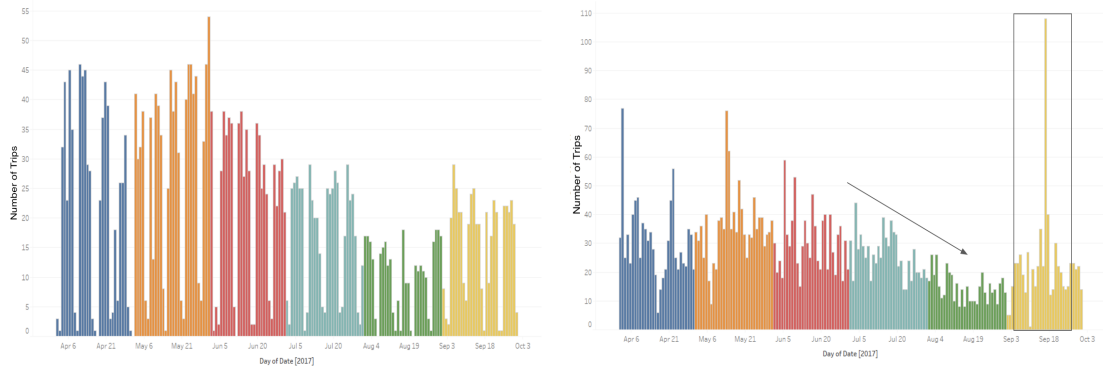
months protests and strikes did not impact bike usage.

Film festival and movie premieres during summer are quite popular in Bologna and gather big number of spectators, leading to a hypothesis that those places might also become attractive for bike users. Thus, we decided to separately look into number of bike rides starting from and ending at two locations which are connected with such events. Figure 18 shows those locations: Figure 18a shows Bologna Fiere[10] which is a large exhibition center with multiple halls, Figure 18b shows Piazza Re Enzo, where many movie premieres happen.

When looking into trips ended in the mentioned places we can see couple of anomalies. In Figure 19a we can see that there was only one day at the end of May that stands out from the other days. When analysing the events that happened on that day we could not find any significantly large event that could have triggered such high number of trips, leading to believe that this was caused with some other reason.

Figure 19b illustrates number of rides ended in the city center near Piazza Re Enzo. Similarly we can see that there are three days that stand out from others with the extremely high number of rides. When looking closely into hourly graphs we noticed that over half of the trips during these days were done in the morning around 7am. Here as well, there were no large event which could have triggered the number of rides. In summary, we can conclude that during the six month period of observation events like movie premieres or large social gathering had not have significant effect on the number of bike rides.

[10]Bologna Fiere - Bologna Exhibitions and Fairs [website] http://www.bolognafiere.it/en/home



(a) Weekly split of the rides that ended in Bologna Fiere

(b) Weekly split of the rides that ended in Piazza Re Enzo

Figure 19: Weekly split of the rides that ended near the mentioned trip locations

5 Predictive Analysis

In this section, we build on insights from previous sections to predict the amount of rides done during one hour and 30 minute time slots. First, we will present the machine learning models that we have used for predicting the number of trips. Secondly, we will present the metrics that were used to evaluate the accuracy of the predictions. Furthermore, we will present the features that are used in the models and the setup of the experiment. Finally, we will present the results from our prediction models for predicting number of trips for 30-minute and 60-minute intervals.

5.1 Machine learning models

Linear regression. Linear regression is one of the most commonly used models for predictive analysis. It estimates the relationship between a dependent variable and independent variable. The number of the independent variable can vary and in case there is more than one independent variable the model is called multiple linear regression. [33] The mathematical formula for the multiple linear regression is the following:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_\mathrm{n} \mathrm{X}_\mathrm{n} + \varepsilon \tag{1}$$

where,

Y is dependent variable,

X are independent (explanatory) variables,

β_0 is Y intercept,

β_p is the slope for each of the independent variable,

ε is the error term or residual.

Random forest. In order to understand random forest model let's look into decision trees. Decision trees are built, as the name suggests, in the form of a tree. It can be used for continuous and categorical output variables. Each internal node in the decision trees indicate a test of an attribute and each branch is the outcome of the test. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance [34]. While Random forest is a supervised learning algorithm, which combines a number of decision trees with some modifications [35]. In short it does not rely only on a single decision tree but uses multiple ones to predict the final value. It is a technique that can be used on both regression and classification tasks.

XGBoost. XGBoost is short from "Extreme Gradient Boosting" which originates from the paper named "Gradient Boosting" written in 2001 by Friedman [36]. Gradient Boosting is a machine learning algorithm which can be used for both regression and classification tasks. It creates a model mostly in the form of ensemble of decision trees. XGBoost is a supervised learning algorithm as well where we try to predict target variable with a training data, which consists of multiple features. It is specifically created to optimize speed and performance of Gradient Boosting models. We should also mention that in 2014 it was reported as one of the best and powerful models in Kaggle Bike Sharing competition [37].

Long short-term memory. In transportation domain one of the most used artificial intelligence algorithms are the convolutional artificial neural network (ANN) and recurrent neural network (RNN). Especially, the later can produce a good predictions on time-series data [38], [39]. However, both RNN and ANN are not working well with the time-series data that have a long time lags. Thus, we have used long short-term memory neural network (LSTM NN) for prediction in our paper as it has the capability of learning long-term dependencies [40], [41]. LSTM are basically designed to avoid problems with long-term dependencies and are now widely used in many areas. RNNs are recently developed deep learning approach which contain a repeating modules of neural networks. While in the standard RNN repeating module usually consists of a single layer, in LSTM the structure is a bit more explicit, it has four modules which are interacting in a special way. In the Figure 20 it can be seen the 4-layer module of LSTM.

A usual LSTM model includes various memory blocks which are called cells (in the above figure cells are the yellow rectangles). Cells have two different states that are transferred from one to another, these states are cell state and hidden state. Besides memory blocks there are tree large mechanisms called gates which are manipulating the memory. Gates are called input gate, output gate and forget gate.

LSTM model is trained using optimization algorithm called 'Adam', which is an ef-

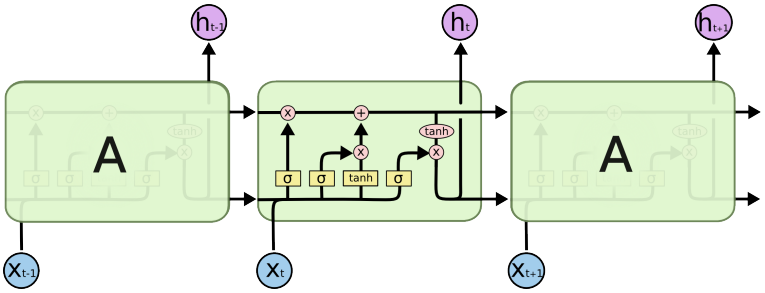


Figure 20: The 4-layer module in LSTM. Source: Christopher Olah,[2015] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

fective method for stochastic optimization [42]. The dropout regularization was used in the model to make sure that we will not get overfitting. We have experimented with different epochs in order to get the best results, additionally we have tried early stopping.

The LSTM model was developed to predict the number of bike rides. The process starts with scaling the historical data, after the model predicted the number of rides based on several variables and historical bike usage data, we unscaled the data and then we calculated Mean Absolute Error (MAE), Mean Squared Error (MSE) , R squared and Root Mean Squared Error (RMSE) in order to compare the predicted results with the actual ones.

5.2 Evaluation of the models

Metrics. For understanding the accuracy and the correctness of the prediction different metrics were used. As the problem we are solving is a regression problem, we are trying to predict a certain number we have used the following metrics to measure our models and choose the best performing model: Mean Absolute Error (MAE), Mean Squared Error (MSE) , R squared and Root Mean Squared Error (RMSE) metrics. Will discuss each metric and how it is calculated below.

MAE - In order to understand MAE, first we need to look a term of Absolute error, which is the amount of errors in our measurements, in other words it is the difference between the observed value (x_0) and the true value (x). The formula for the absolute error is quite straightforward:

$$\text{Absolute Error} = x_0 - x \tag{2}$$

MAE is described as a the average of all absolute errors. Formula looks the following way:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \tag{3}$$

MSE is the average of the squared of all absolute errors. The squaring is done in order to get rid of negative signs. It also gives larger differences in case of outliers. Formula is as follows

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \tag{4}$$

RMSE is the standard deviation of residuals. Residuals are the difference between the actual value and the value that our model predicted. Formula is the following:

$$RMSE = \sqrt{(f - o)^2} \tag{5}$$

R squared is measure which shows what portion of the dependant variable can be explained by an independent variable. The formula:

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \tag{6}$$

5.3 Features used for the models

For the predictions we decided to use the following features:

1. Weekday: We have noticed that there is a tendency in riding bikes more during weekdays. In contrast to weekends when the usage of bikes is dropping heavily.

2. Hour: We found that the usage of bikes varies across different points of the day, with certain time slots having significantly more rides.

3. Month: In Section 4.1 we noticed that there are more rides done during May and only few rides during August, when most of the people were out of the town. The number of rides started to increase during September when people returned back to Bologna.

4. Season: Similarly, we can see a tendency during different seasons, even though we did not have sufficient data to analyse the bike usage during Autumn and Winter, we still can conclude that number of bike usages is dropping during Summer and the highest number of rides are done in Spring.

5. Temperature: In Section 4.4 we noticed that that higher temperature means less rides and lower temperature leads to more bike usages.

6. Holiday: In Chapter 4.6 we could see that not during all holidays there is a significant difference in number of rides, but in some cases it does cause a significant drop, that is why we decided to use the feature in the models.

7. New features: Lastly, we have created new variables from our existing features such as number of rides during one hour ago and number of rides one week ago.

We have decided not to use features such as Wind speed and Air Pollution value in our final experiments as we have noticed from Sections 4.4 and 4.5 that these two features did not affect the number of rides significantly, and will not help our model in predicting accurate numbers.

5.4 Experiment Setup

We aim to predict number of rides done during the next 30-minute and 60-minute intervals, by splitting the data into training and testing sets.

We have prepared two datasets. In one we have aggregated number of trips by hours and in the second we aggregated number of trips by 30 minute time slots. In both datasets we have removed outliers and converted categorical variables into dummy variables, this way our machine learning models will present a better results. As an example for categorical variable 'Weekday' we have created 7 different columns with the names of a weekday with 1s and 0s.

Next, we created correlation heat map, which helped us with identifying incidence pattern and discover anomalies. On Figure 21 we can see that number of rides are mostly correlated with 'Temperature' variable, 'Month' variable and 'season_2' dummy variable.

We have experimented with our model by adding new variables 'Count-1', which is number of rides done one hour ago and 'Week-1' which is number of rides done one week ago on the during the same time period, as we have noticed in Section 4.6 there is a strong weekly trend.

As mentioned before we have experimented with several prediction models, such as Linear Regression, Random Forrest Regression, XGBoost Regression and, compared the results with the Long-Short Term Memory(LSTM). LSTM gave the best results in terms of RMSE, MAE, MSE.

We started with splitting data into 80% into training and 20% into test splits, which is one of the most commonly accrued ratios [43]. Later, we have experimented with different split ratios and found out that 90/10 ratio gives the best results.

K Fold Cross Validation. In order to improve the results from predictions done by LSTM model we have also used K fold Cross Validation with 5K and 10K. In statistics cross validation is a method which is used to evaluate machine leaning models by resampling the data [44]. In K fold Cross Validation, K stands for a parameter which shows how many groups the given data set will be split into. For our experiment we have splitted our

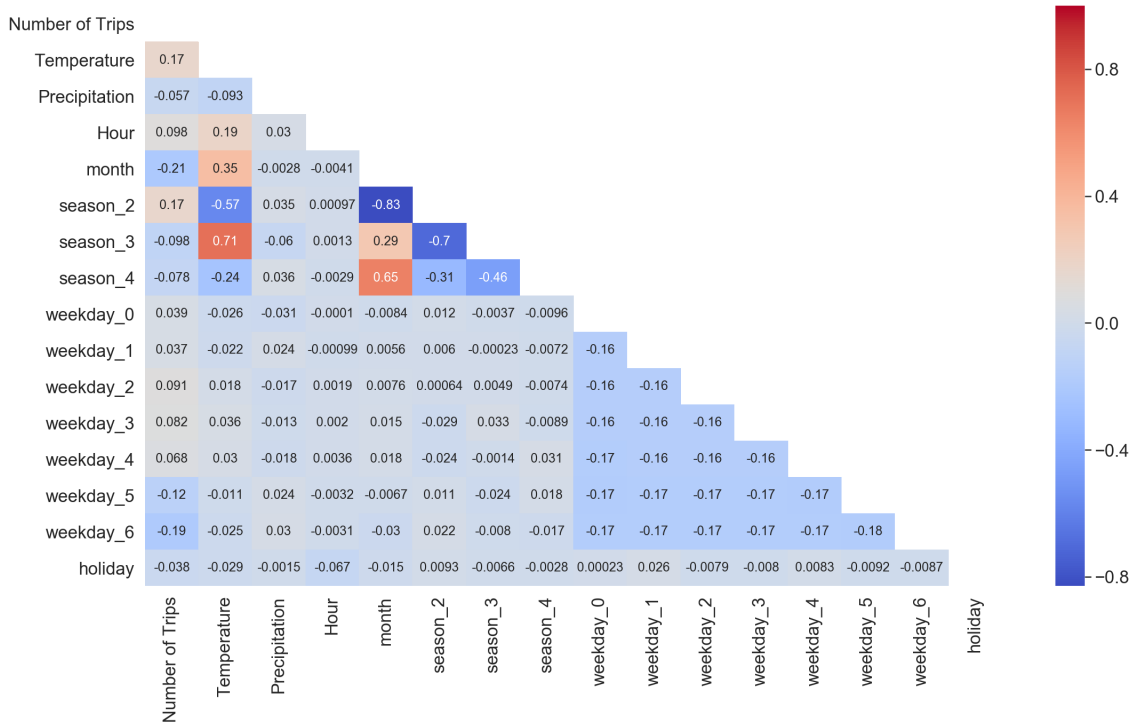


Figure 21: Correlation Heat Map of Variables fro Bella Mossa data. Source: Author's calculations.

dataset into 5 equal sizes and later into 10 equal sizes. Figure 22 shows the process of K fold Cross Validation with 5 iterations.

We have experimented with 5-fold Cross Validation and 10-fold Cross Validation.

5.5 Prediction results

We have conducted numerous experiments using different features, but for illustration purposes we have decided to include the best ones.

Figure 23 illustrates the results that we have received when predicted the number of rides and compering them with the actual amount of the rides. Each time step on the graph is 30 minute. We can see that the model predicts the number of rides very accurately and the prediction line is accurately repeats the true(actual) number of rides.

We have reached this by using features mentioned in subsection 5.4. We noticed that we get the best results in such way, additionally we have used K fold Cross Validation with 10K and with 5K, the former giving much lower error rate. On Figure 3 we report these results.

In Table 3 we provide results from different models, we have used 90-10 split for these experiments and tried predicting number of rides during 60-minute interval. We can clearly see that LSTM is almost 4.5 times better than other models.

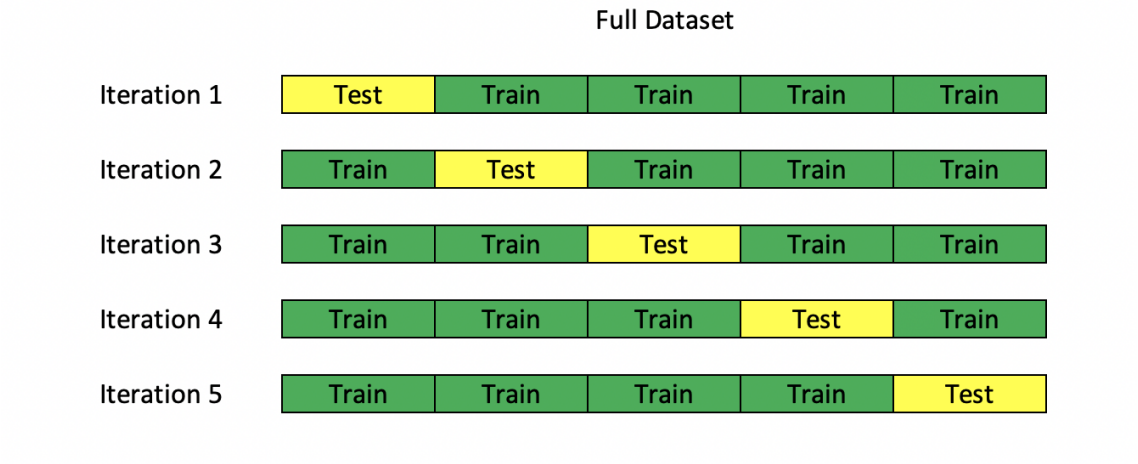


Figure 22: Process of K fold Cross validation with 5 iterations.

Name of the Metric	LSTM	Linear Regression	Random Forrest	XGBoost
MAE	16.62	46.03	44.72	47.04
MSE	457.73	4438.29	4354.25	4809.13
RMSE	21.39	66.62	65.98	69.34
R-squared	0.91	0.25	0.33	0.17

Table 3: Prediction statistics for different models, 90-10 split, 60 minute interval

Second best model for predicting number of bike usage is Random Forrest which is slightly better than Linear Regression and XGBoost. Table 5 we compare the results of experiments with different data split. The best results were received with 90/10 split ratio. Second best was 70/30 ratio which was just slightly worse than 90/10. 80/20 showed the biggest error out of the three splits.

Metrics	90/10 split	80/20 split	70/30 split	60/40 split
MAE	16.62	23.62	25.44	16.75
MSE	457.73	986.56	1140.87	546.67
RMSE	21.39	31.40	32.45	23.38
R-squared	0.91	0.63	0.70	0.75

Table 4: Prediction statistics for LSTM model for 60-minute interval with different split ratios.

Seeing such results we have decided to focus more on LSTM model and tried to improved the error rates. First, we have experimented with adding new variables such

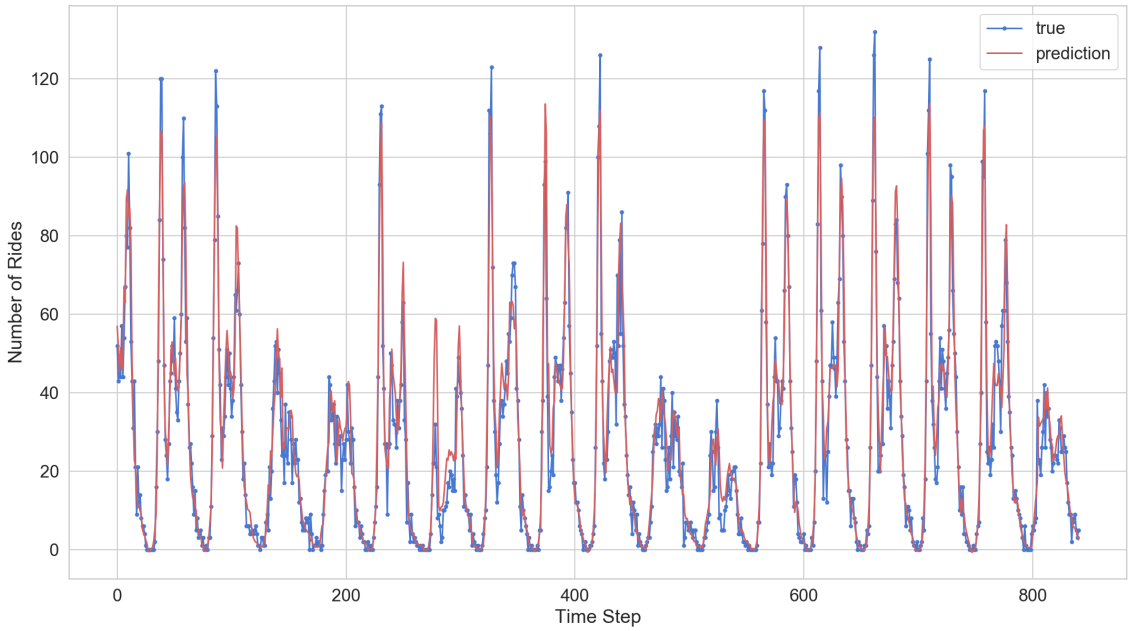


Figure 23: Correlation Heat Map of Variables fro Bella Mossa data. Source: Author's calculations.

as 'Count-1' (trips done one hour before) and 'Week-1' (trips done exactly one week before). We could not achieve the results we have been hoping for even though we have seen positive change in the results. When using 'Count-1' feature we have noticed that our model has improved by 30% when looking into MSE, MAE and RMSE results. Feature 'Week-1' improved our model furthermore by around 45% from the initial experiment, we have reached MAE of 11.21. Both results are received when using 90-10 data split.

Furthermore, we have used our model to predict number of rides for 30 minute intervals. And we have seen significant improvement in the metrics. In Table 5 we report the results using different data splits. The figure shows that when predicting with 30-minute interval we improve the model over 2 times in comparison with 60-minute predictions. We have reached the best results using K fold Cross Validation with 10K-s.

Metrics	90/10 split	80/20 split	70/30 split
MAE	5.38	6.42	5.62
MSE	66.08	104.54	70.63
RMSE	8.12	10.22	8.40
R-squared	0.91	0.86	0.87

Table 5: Prediction statistics for LSTM model for 30-minute interval with different split ratios

Finally, we can conclude that LSTM gives relatively accurate results in terms of forecasting number of rides for 30 minute and 60 minute intervals. We get much better results with LSTM in comparison with other machine learning algorithms and this can be improved even more in case we have a larger dataset.

6 Conclusions and Future Work

We conduct and present behavioral analysis of bike usage in the city of Bologna. We look into temporal patterns and spatial patterns, furthermore, we analyze the impact of weather conditions such as air temperature, amount of rains, and wind speed. The analysis is performed with a data set received from the Bella Mossa Experiment, which consisted of several transportation means and over 320,000 bicycle rides during the period from April 2017 till the end of September 2017. Finally, we build several models to predict the number of rides for a specific period.

In the analysis part, we find that the most popular month for bike riding in May and the least popular month is August, with summer being the least popular season for cycling in Bologna. We find that 83% of total rides are done during weekdays and only 17% are done on Saturday and Sunday. This pattern is not changing throughout different months. We noticed that on average most of the rides are done during mid-week, on Wednesday and Thursday and the fewest number of rides are done during Sunday. Furthermore, most of the rides are done at 6 am and 5 pm on weekdays, which is mainly connected with people going to work or school and taking a return ride. Finally, during weekends we notice that many rides are happening close to midnight indicating that many people go out and enjoy the nightlife.

In the spatial analysis part, we find that the most attractive points in the city for bike riding are City Center, Train Station, and the area around Bologna University, near Porta Saragozza. 70% of the total rides are done in these three hubs. The pattern is changing during Summer months when Universities are closed and the number of rides around that area decreases heavily. Two out of three busiest streets are the main streets of Bologna, Via Rizzoli and Via dell' Indipendenza, located in the city center, these two streets contain most of the rides during the whole period of the experiment. We find also that during evening busiest streets from the city center are moving into streets leaving the city center such as Via Sabotino and Via Mazzini. This implies the fact that many people return to their homes which are located outside the city center.

Next, we analyzed how people spread out from the most attractive hubs and found that from Train station bike users go to the City Center, to the University of Bologna Department of Computer Science and Engineering, and the Ospedale Maggiore "Carlo Alberto Pizzardi". We notice that this changes heavily during August in comparison to May. During August users mainly took a path to the city center, very few rides were done from Train Station to University buildings or the Ospedale Maggiore. From City Center during May most bikes are spreading into residential neighborhoods of Porta Saragozza and Murri, indicating that many people return to their homes after spending some time in

the city center. This pattern stays almost the same throughout the six 6 months period. From the University of Bologna facilities, users mainly take bike trips to residential areas and areas near Porta San Felice. This patterns is the same during spring months but changes heavily in summer as fewer and fewer users take trips to the University buildings.

Besides, we find that number of bike riders is changing due to weather conditions such as air temperature and amount of precipitation. We find that there is a negative correlation between the number of rains and the number of bike rides. We also notice that higher temperature leads to a larger number of rides to some extent and when the temperature crosses a certain level (around 28°Celsius) number of rides start to drop. Regarding the correlation between air pollution and bike usage, we did not find any proof that bike usage during the six months has significantly decreased the pollution level. Finally, we found that during some holidays bike usage is significantly lower.

In the prediction part, we predict the number of rides done at a given time. We find that the best prediction results are received using Neural Networks, in particular the LSTM model. We noticed that we get better results when splitting data into 30-minute bins and predicting the number of rides done during the next 30 minutes. We perform K fold Cross-Validation with 5Ks and 10Ks and find that the later gives better results. Finally, we report that the prediction results are significantly strong in our models giving MAE of 5.3.

For future work, several directions can be considered. First, our dataset consisted of only six months of data, this was not letting us analyze several seasonal factors while having a larger dataset with a longer timeline might give more insights in the bike usage analysis and might improve the prediction results further. Secondly, the patterns of bike users can be compared with the patterns of users from different cities. Additionally, public transport data can be analyzed to understand how trains, buses, and taxis affect bike usage and how this pattern changes throughout a year. Finally, other RNN models such as Grated Recurrent Unit (GRU) can be used to predict the number of rides.

In conclusion, based on the results we received we can indicate that Neural Networks and, LSTM model, in particular, have quite a high accuracy when predicting the number of trips during different time frames. Our model gives better results than more frequently used machine learning and statistical models and can be used on numerous occasions to predict the future number of rides. The prediction model can be used when analyzing which streets might be utilized and which will lack bike rides and will not be used as much as intended. Additionally, the model can be used to predict the demand gap, which later used for redistributing bikes in the city. And making sure that people who would like to use bikes will most certainly find one near their location.

References

[1] Thomas Nosal and Luis F Miranda-Moreno. The effect of weather on the use of north american bicycle facilities: A multi-city analysis using automatic counts. *Transportation research part A: policy and practice*, 66:213–225, 2014.

[2] Wafic El-Assi, Mohamed Salah Mahmoud, and Khandker Nurul Habib. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in toronto. *Transportation*, 44(3):589–613, 2017.

[3] Max Nankervis. The effect of weather and climate on bicycle commuting. *Transportation Research Part A: Policy and Practice*, 33(6):417–431, 1999.

[4] Jon Edward Froehlich, Joachim Neumann, and Nuria Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[5] Tongtong Liu, Zheng Yang, Yi Zhao, Chenshu Wu, Zimu Zhou, and Yunhao Liu. Temporal understanding of human mobility: A multi-time scale analysis. *PloS one*, 13(11):e0207697, 2018.

[6] Stephen J Mooney, Kate Hosford, Bill Howe, An Yan, Meghan Winters, Alon Bassok, and Jana A Hirsch. Freedom from the station: Spatial equity in access to dockless bike share. *Journal of transport geography*, 74:91–96, 2019.

[7] Grant McKenzie. Docked vs. Dockless Bike-sharing: Contrasting Spatiotemporal Patterns (Short Paper). In Stephan Winter, Amy Griffin, and Monika Sester, editors, *10th International Conference on Geographic Information Science (GIScience 2018)*, volume 114 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 46:1–46:7, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[8] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2015.

[9] Jiawei Zhang, Xiao Pan, Moyin Li, and S Yu Philip. Bicycle-sharing system analysis and trip prediction. In *2016 17th IEEE international conference on mobile data management (MDM)*, volume 1, pages 174–179. IEEE, 2016.

[10] Nguyen Duc-Nghiem, Nguyen Hoang-Tung, Aya Kojima, and Hisashi Kubota. Modeling cyclists' facility choice and its application in bike lane usage forecasting. *IATSS research*, 42(2):86–95, 2018.

[11] Cheng Zhang, Linan Zhang, Yangdong Liu, and Xiaoguang Yang. Short-term prediction of bike-sharing usage considering public transport: A lstm approach. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1564–1571. IEEE, 2018.

[12] Chengcheng Xu, Junyi Ji, and Pan Liu. 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, 2018.

[13] Yi Ai, Zongping Li, Mi Gan, Yunpeng Zhang, Daben Yu, Wei Chen, and Yanni Ju. A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system. *Neural Computing and Applications*, 31(5):1665–1677, 2019.

[14] Brian Caulfield, Margaret O'Mahony, William Brazil, and Peter Weldon. Examining usage patterns of a bike-sharing scheme in a medium sized city. *Transportation research part A: policy and practice*, 100:152–161, 2017.

[15] Susan L Handy and Yan Xing. Factors correlated with bicycle commuting: A study in six small us cities. *International Journal of Sustainable Transportation*, 5(2):91–110, 2011.

[16] Justine Sears, Brian S Flynn, Lisa Aultman-Hall, and Greg S Dana. To bike or not to bike: Seasonal factors for bicycle commuting. *Transportation research record*, 2314(1):105–111, 2012.

[17] Elliot Fishman, Simon Washington, Narelle Haworth, and Angela Watson. Factors influencing bike share membership: An analysis of melbourne and brisbane. *Transportation research part A: policy and practice*, 71:17–30, 2015.

[18] Javier Molina-García, Isabel Castillo, Ana Queralt, and James F Sallis. Bicycling to university: evaluation of a bicycle-sharing program in spain. *Health promotion international*, 30(2):350–358, 2015.

[19] Daniel Fuller, Lise Gauvin, Yan Kestens, Mark Daniel, Michel Fournier, Patrick Morency, and Louis Drouin. Use of a new public bicycle share program in montreal, canada. *American journal of preventive medicine*, 41(1):80–83, 2011.

[20] Ole Hertel, Martin Hvidberg, Matthias Ketzel, Lars Storm, and Lizzi Stausgaard. A proper choice of route significantly reduces air pollution exposure—a study on bicycle and bus trips in urban streets. *Science of the total environment*, 389(1):58–70, 2008.

[21] Christer Johansson, Boel Lövenheim, Peter Schantz, Lina Wahlgren, Peter Almström, Anders Markstedt, Magnus Strömgren, Bertil Forsberg, and Johan Nilsson Sommar. Impacts on air pollution and health by changing commuting from car to bicycle. *Science of the total environment*, 584:55–63, 2017.

[22] Michael Chertok, Alexander Voukelatos, Vicky Sheppeard, and Chris Rissel. Comparison of air pollution exposure for five commuting modes in sydney-car, train, bus, bicycle and walking. *Health promotion journal of Australia*, 15(1):63–67, 2004.

[23] Jillian Strauss, Luis Miranda-Moreno, Dan Crouse, Mark S Goldberg, Nancy A Ross, and Marianne Hatzopoulou. Investigating the link between cyclist volumes and air pollution along bicycle facilities in a dense urban core. *Transportation Research Part D: Transport and Environment*, 17(8):619–625, 2012.

[24] Jinbao Zhao, Jian Wang, and Wei Deng. Exploring bikesharing travel time and trip chain by gender and day of the week. *Transportation Research Part C: Emerging Technologies*, 58:251–264, 2015.

[25] Pierre Borgnat, Patrice Abry, Patrick Flandrin, Céline Robardet, Jean-Baptiste Rouquier, and Eric Fleury. Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(03):415–438, 2011.

[26] Haitao Xu, Jing Ying, Hao Wu, and Fei Lin. Public bicycle traffic flow prediction based on a hybrid model. *Applied Mathematics & Information Sciences*, 7(2):667, 2013.

[27] Divya Singhvi, Somya Singhvi, Peter I Frazier, Shane G Henderson, Eoin O'Mahony, David B Shmoys, and Dawn B Woodard. Predicting bike usage for new york city's bike sharing system. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.

[28] Susan Shaheen and Adam Cohen. Shared micromoblity policy toolkit: Docked and dockless bike and scooter sharing. 2019.

[29] Hao Luo, Zhaoyu Kou, Fu Zhao, and Hua Cai. Comparative life cycle assessment of station-based and dock-less bike sharing systems. *Resources, Conservation and Recycling*, 146:180–189, 2019.

[30] Tianqi Gu, Inhi Kim, and Graham Currie. To be or not to be dockless: Empirical analysis of dockless bikeshare development in china. *Transportation Research Part A: Policy and Practice*, 119:122–147, 2019.

[31] Yu Shen, Xiaohu Zhang, and Jinhua Zhao. Understanding the usage of dockless bike sharing in singapore. *International Journal of Sustainable Transportation*, 12(9):686–700, 2018.

[32] Jonathan Corcoran, Tiebei Li, David Rohde, Elin Charles-Edwards, and Derlie Mateo-Babiano. Spatio-temporal patterns of a public bicycle sharing program: the effect of weather and calendar events. *Journal of Transport Geography*, 41:292–305, 2014.

[33] T L_etc Lai, Herbert Robbins, and C Zi Wei. Strong consistency of least squares estimates in multiple regression ii. *Journal of Multivariate Analysis*, 9(3):343–361, 1979.

[34] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[35] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[36] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[37] Kaggle. Kaggle, 2015. bike sharing demand [www document]. "`https://www.kaggle.com/c/bike-sharing-demand`", (accessed: 06.30.2020).

[38] Jia-Shu Zhang and Xian-Ci Xiao. Predicting chaotic time series using recurrent neural network. *Chinese Physics Letters*, 17(2):88, 2000.

[39] Jun Zhang and KF Man. Time series prediction using rnn in multi-dimension embedding phase space. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, volume 2, pages 1868–1873. IEEE, 1998.

[40] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[41] Zheng Zhao, Weihai Chen, Xingming Wu, Peter CY Chen, and Jingmeng Liu. Lstm network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2):68–75, 2017.

[42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[43] George EP Box and R Daniel Meyer. An analysis for unreplicated fractional factorials. *Technometrics*, 28(1):11–18, 1986.

[44] Frederick Mosteller and John W Tukey. Data analysis, including statistics. *Handbook of social psychology*, 2:80–203, 1968.

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Taron Davtyan**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Forecasting Bike Demand. Bologna Case Study,

   supervised by Rajesh Sharma and Flavio Bertini.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Taron Davtyan
08/08/2020