Smart Public Transport System:London's Bicycle Sharing Scheme

Juan Navarrete

THE UNIVERSITY
of EDINBURGH

Master of Science

Artificial Intelligence

School of Informatics

University of Edinburgh

2016

Abstract

The surge in travel demand caused by urbanization is making it increasingly difficult for transportation service operators to keep the expected high quality of service. The use of science and technology to create Smart Transportion Systems can increase the efficiency of the existing urban infrastructure to mitigate this issue. This MSc project aims to make London's bicycle sharing scheme, Santander Cycles, smarter by discovering valuable insights about the operational and usage patterns and by predicting bicycle availability 10, 60 and 180 minutes ahead in each of it's stations. This increased understanding of the system will help devise data-driven strategies that make the bicycle sharing scheme more efficient, reliable and pleasant to use.

Acknowledgements

I would like to thank my project supervisors Nigel Goddard and Pedro Baiz for their help and guidance in the development of this project, my family for their unconditional love and support, and the Mexican Council of Science and Technology for providing me the financial support to pursue my postgraduate studies.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Juan Navarrete*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Challenges of Urbanization

The world has been experiencing unprecedented levels of urban growth in the past decades. For the first time in history, more than half of the world's population (54 %) lives in cities, and by 2050, projections suggest this number will increase to 66% [20]. The city of London is not exempt from this trend. The Greater London Authority forecasts that the city's population will reach 10.4 million by 2041, an increase of 22.3 % from today's measurements [2]. While this phenomenon, called urbanization, has the potential to bring important cultural, economic, and societal benefits, it also causes a surge in the demand of public services such as transportation, education, housing, electricity and water, which if not managed adequately, could overwhelm the city's infrastructure and exacerbate current problems or create new ones. Therefore, it is particularly important that governments, companies and civil society plan and adapt accordingly to mitigate the threat that urbanization poses to the sustainability of the city.

Traditionally, increased demand has been mitigated by expanding the city's infrastructure, e.g. building new roads, laying additional pipelines, acquiring new buses, etc. However, this approach is not sustainable, as it is not environmentally friendly nor cost-effective. Furthermore, it depends upon the availability of large free spaces, which are finite in urban areas [8]. Thus, as an effort to keep pace with the city's growth and to close the infrastructure gap, the research community has centered it's attention on the use of science and technology to increase the efficiency of the existing urban infrastructure. This has led to the development of "smart" cities, which via advanced applications such as Intelligent Transport Systems (e.g. traffic monitoring,

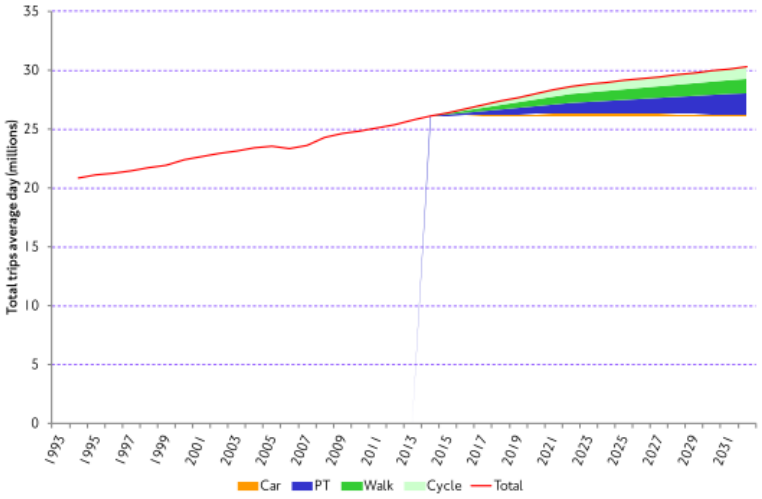

Figure 1.1: Projected trip growth to 2031, 'Low Car' scenario (conserative outlook in fuel efficiency improvements and reduced car ownership). Adapted from Travel in London Report 8, Mayor of London, 2015 [22]

asset management and smart parking), Assisted Living (e.g. telehealth and telecare systems), Water Management Systems (e.g. consumption monitoring and flood management), Smart Grids (energy efficiency programs and renewable energy integration) and Waste Management Systems (waste collection modeling and consistent supply to energy generation programs), aim to improve the quality of life of the cities' residents [27]. On this project we focus on Intelligent Transport Systems (ITS) and their use in the city of London.

1.2 Intelligent Transport Systems

Despite the well spread and developed transportation network of London, increased urbanization introduces mobility and environmental problems due to saturation of the infrastructure. For example, on average, the typical Londoner makes 2.5 trips daily, so, by 2041, 5.5 million additional trips will be made every day through the transportation network of the city [22]. This surge in demand makes it increasingly difficult for service operators to guarantee the high quality of service that users expect. Implementing technological solutions, which provide users better decision support and assist authorities to manage investments, assets and information, is becoming a necessity to improve mobility in the city. ITS describe the use of a broad range of disciplines such as information processing, control, sensors and advanced communications, to create

smart users and infrastructure which improve the environmental performance, safety and efficiency of the transportation network of a city [6]. Smart Public Transport Systems (SPTS) are the type of ITS that focus on public transportation and ride sharing schemes [4]. In the case of London, these include the Tube, boats, buses, trains and Santander Cycles, the city's bicycle sharing scheme.

1.3 Cycling in London

One of the key goals of recent London's administrations has been to make cycling an integral part of the transportation network of the city, as it not only has the potential to improve the residents' quality of life by providing a healthy and cost-effective way of transport, but also to improve urban sustainability by taking petrol-powered vehicles off the streets and freeing spaces in trains, boats and buses. [21]. Major investments and strategic policies such as building bicycle superhighways and bicycle awareness programs have been made to increase cycling in the city. However, according to a report developed for the Central London Air Quality Cluster Group [24], one of the most economical ways to achieve this is through the city's bicycle sharing scheme, as for example, the cost per additional cyclist of cycle lanes is between £2,600 and £98,000 while for Santander Cycles it is between £440 and £990. With the increased role of the bicycle sharing scheme for the success and sustainability of the city's transportation network, maintaining high operational levels and providing a good customer experience is becoming particularly important.

This project focuses on Santander Cycles, which consists of more than 11,000 bicycles distributed in 779 stations throughout the city of London. The service, owned by Transport for London (TfL) but operated by Serco, allows it's members (presently 208,957) and casual users to borrow and return bicycles from any docking station in the scheme on a short term basis, 24/7, 365 days a year.

While widely accepted and used, the bicycle sharing scheme is not without problems. The latest Cycle Hire Customer Satisfaction and Usage Survey [10] shows the lack of docking space and bicycle availability as both the most frequent reason for users not renewing their membership and the most mentioned answer when asked about bad experiences with the service. Furthermore, tweets mentioning this same problem, such as Pascual V. Daz (yanovengomas). "Another morning, same story. REALLY BAD SERVICE of @SantanderCycles in Moorgate area. No empty docks 1 mile around.". 07 Jul 2016, 09:16 UTC. Tweet, are not rare among @SantanderCycles (the official bi-

Figure 1.2: A Santander Cycles docking station in Southwark. Nick-D (User). (2015) Santander Cycles docking station in Southwark during November 2015 [Digital image]. Retrieved from Wikimedia Commons website: `https://upload.wikimedia.org/wikipedia/commons/2/27/Santander_Cycles_docking_station_in_Southwark_during_November_2015.jpg`

cycle sharing scheme Twitter account) mentions. Though the service provider has implemented measures to mitigate the availability problems, e.g. deploying teams which redistribute the bicycles among stations using trucks, TfL recognizes that it remains the key area of improvement in the service [10].

1.4 Project Goals

The goal of this project is to use data mining and machine learning to improve the service levels of London's Santander Cycles sharing scheme by:

1. Achieving a better understanding of the sharing scheme's behavior by finding valuable information about it's operation and usage patterns

2. Developing a system which predicts the number of bicycles available in each docking station. With such system, the service operator would be able to know in advance when bicycle redistribution is needed in order to prevent the stations from becoming full or empty. Furthermore, if made public, the forecasts would allow users of the scheme to plan their journey better to avoid such troublesome stations.

With these developments, we aim to enable the implementation of data-driven strategies which make Santander Cycles more pleasant to use and therefore, more widely adopted. This, in turn, would improve Londoners' quality of life by helping establish a more sustainable transportation network.

Chapter 2

Background

2.1 Bicycle Sharing Schemes

A bicycle sharing scheme is a service in which bicycles are made available via docking stations distributed throughout a city for their shared use by individuals on a short term basis [18]. By allowing users to borrow a bike from docking station A and return it to docking station B, each bicycle can be used by several users each day as a healthy, cheap and sustainable mean of transport. This type of systems was first introduced in Amsterdam in the year of 1965, but it was thanks to advancements in information technology such as user's databases, journey tracking, GPS, smartcards, and real-time availability mobile applications, that they became viable and widely popular. As of today, more than 1,000 bicycle sharing schemes are operational throughout the world (Figure 2.1), making available for hire approximately 1,392,170 bicycles in cities located in America, Europe, Asia, Australia and Africa [17].

Figure 2.1: Bicycle sharing schemes throughout the world. (2016) The Bike-sharing World Map [Digital Map]. Retrieved from website: `http://www.bikesharingmap.com`

We use the following metrics [12] to describe a docking station's characteristics, status and behavior at a given moment of time t or time period p:

- **Size (SZ):** Measures the size of a station at time t.

$$SZ(t) = \alpha_t + \beta_t \tag{2.1}$$

  where α_t and β_t are the number of available bicycles and the number of available empty docking spaces at time t, respectively. Can vary due to docking spaces or bicycles being not functional.

- **Activity Score (AS):** Measures how active a station is at a given time.

$$AS(t) = |\alpha_t - \alpha_{t-1}| \tag{2.2}$$

  where α_t is the number of bicycles at time t and α_{t-1} the number of bicycles in the immediate previous reading.

- **Normalized Activity Score (NAS):** AS divided by the station size to disregard the varying stations sizes.

$$NAS(t) = \frac{AS(t)}{SZ(t)} \tag{2.3}$$

- **Binary Activity Score (BAS):** Binary version of AS, where $ES(t) = 1$ if $AS(t) > 0$, else $ES(t) = 0$.

- **Normalized Available Bicycles (NAB):** Measures the fullness of a station at a given time.

$$NAB(t) = \frac{|\alpha_t|}{SZ(t)} \tag{2.4}$$

  where α_t is the number of available bicycles a time t.

- **Readings Count (RC):** Measures how active a station is as determined by the number of readings that occurred during time period p. Assumes that readings are only published whenever the number of bicycles in a station changed.

- **DayView:** Reflects a station's behavior throughout a 24 hour period discretized in 288 five-minute bins as determined by one of the previously defined metrics. For example, the NAS Weekend Dayview is obtained by averaging the NAS values of each of the 288 bins across every weekend day during our observational period.

- **Priority:** Metric that defines how important a station is. Can have 3 values: 1 (High Priority), 2 (Normal Priority) and 3 (Priority not Available).
- **Distance Metric:** Metric used to measure the similarity between two time series of a station's data. As suggested by [12], we use Dynamic Time Warping with a one-hour Sakoe-Chiba band as it handles temporal shifts of up to one hour which might occur due to seasonality, unexpected events or other reasons.
- **Empty Station:** A station with no fully functional bicycles is considered to be empty. One or more bicycles marked for repair can be present in an empty docking station.
- **Full Station:** A station with no fully functional free docking spaces is considered to be full. One or more docking spaces marked for repair can be present in a full docking station.

2.2 Related Works

Other studies analyzing and predicting the behavior of bicycle sharing schemes throughout the world have been done before by talented scholars:

First, Froehlich et al [12] used clustering to identify stations with common operational patterns and a Bayesian Network (BN) to predict bicycle availability 5 and 120 minutes ahead. They used data collected during 13 weeks from the 390 docking stations of Bicing, Barcelona's sharing scheme. The researchers found that a station's usage is related to the cultural and geographical characteristics of it's location, which, for example, allowed them to infer if a station's neighborhood was residential or commercial based on it's usage patterns. A limitation of their availability prediction model is that it does not include any exogenous variables into consideration, which makes it miss external influences that affect usage patterns, such as the presence of rain. Kaltenbrunner et al [15] experimented with an Auto-Regressive Moving Average (ARMA) model to predict bicycle availability in this same sharing scheme. But in contrast to the studes done by Froehlich et al, they included in their models variables that represent the number of bicycles available in neighboring stations, which improved the performance of their predictor. This sounds reasonable as a full or empty station forces users to use one of the surrounding stations instead. They concluded that considering 5 to 20 stations increases predictive power (with 15 being the optimal number), but that including less than that or more, worsens it.

Then, research to uncover spatio-temporal patterns and to predict the number of bicycle rentals on a daily basis was made by Borgnat et al [3] using a dataset consisting of two years of records of every journey made in Lyon's bicycle sharing scheme, Velo'v. Their predictor, a Linear Regression (LR) model, was the first one to include non-time related exogenous variables (independent variables that are not part of the system per se but that influence it) such as the amount of rain in millimeters. Although their model was simple, the use of these additional variables improved it's performance and helped it be accurately enough for the coarser grained prediction task that the team had in hand.

Interesting studies have been done targeting Dublin's Dublinbikes as well. First, Yoon et al [32] developed a personal journey advisor to help people move quicker through the city by recommending the pair of stations which minimizes the overall biking and walking travel time. A fundamental component of the advisor is a predictive model which helps suggest only stations that have a high probability of containing available bicycles or free docking spaces. Their model, a time series statistical analysis technique called Auto-Regressive Integrated Moving Average (ARIMA), learns the spatio-temporal usage patterns by using variables representing neighboring stations and the weekly seasonal trend. They experimented with Voronoi Tesselation based, K-nearest Similar pattern based and Linear Regression based approaches to discover relevant surrounding stations for their predictions and concluded that the Linear Regression based approach produced the best results. Furthermore, they determined that while including the influence of the surrounding stations was helpful for short-term predictions, it was detrimental for long-term predictions (60 minutes). Later, Chen et al [5] developed a two stage algorithm to first, predict bicycle availability in the docking stations of this same system and, second, to predict the waiting time for a bicycle to become available in case a docking station was empty. The former was achieved by using a Generalized Additive Model (GAM) along with additional weather exogenous variables such as fog and humidity, while the later by fitting an exponential distribution with time-varying intensity. Their results showed that modeling the non-linear effects of factors such as weather and time of day helps make more accurate medium-term and long-term predictions.

Giot et al [13] then experimented with several machine learning regression approaches such as Ridge (RR), Adaboost, Support Vector, Random Forest and Gradient Tree Boosting regression to predict the number of bicycles in use in the city of Washington up to 24 hours ahead. Given that their dataset did not include data from

individual stations, their predictions have city-wide granularity. They used variables similar to the ones used by Chen et al, which they ranked based on their relevance as determined by the weights of Ridge Regression. Their analysis showed that the number of bicycles in t_{-n} hours and the temperature were the most relevant, that humidity, weekday, holiday and season were averagely relevant and that hour, windspeed and working day were not relevant at all. However, it should be noted that, since the relevance was determined by a linear predictor, a non-relevant variable may be considered like that not because it is really not relevant, but because it's relevance relies on non-linearities. They discovered that flexible state of the art models are highly susceptible to overfit this problem and suggest that special measures to prevent this issue such as L1 regularization should be used to avoid it in future works.

Recently, Yang et al [30] [31] studied the use of Neural Networks (NN) to predict bicycle availability in the docking stations of Bicing 10, 20, 30, 60, 90 and 120 minutes ahead. For such purpose, they collected data during 11 weeks which was then split as follows: the first seven weeks to train the network and the last 4 weeks to assess it's performance. Similarly to other works [32][15], they included the influence of neighboring stations into their model, which again showed to improve the model's performance. However, they did not include non-time exogenous variables such as the weather, which might helped too. They conclude that the number of optimal stations to take into account is 10 and suggest that future work with additional features such as other public transportation options could improve the predictor's performance.

To the best of our knowledge, the only work which targets London's Santander Cycles was carried out by Rylander et al [25], who developed a model based on K-means clustering and Polynomial Regression to first, train a model for each cluster of stations of similar location and popularity, and then, to make predictions of the system's total bicycle usage and for each individual station. They used the cycle hire journey data published by TfL, which includes the time and start/end station of every rental done from 2012 to 2015, and augmented it with weather, seasonal and previous availability variables. In contrast to [13], they state that the use of Ridge Regression (which imposes a regularization term to the model) has no significant impact in the performance and resort to using a 2nd Degree Polynomial (2ND DPOLY) model instead. They suggest that their algorithm was limited by the number of features available and that future work which includes additional variables could be done.

Table 2.1 on page 12 provides a summary of these previous works. Note that a fair comparison between them can not be made due to the lack of a standardized evaluation

metric and the use of different datasets.

While these previously discussed studies have each discovered different features that are relevant to the task of predicting bicycle availability, none of them has actually used a combination of all the variables that have been found as effective. Our research improves the state of knowledge in this field by experimenting with this theoretical optimal combination of features, and, by including bicycle redistribution data in the model, which was ignored before or was simply not available.

Author	Froehlich [12]	Kaltenbrunner[15]	Borgnat[3]	Yoon[32]	Chen [5]	Giot [13]	Yang[31]	Rylander[25]
Model	BN	ARMA	LR	ARIMA	GAM	RR	NN	2ND DPOLY
Variables								
Time	✓	✓	✓	✓	✓	✓	✓	✓
# of Bikes	✓	✓	✓	✓	✓	✓	✓	✓
Autoregressive		✓	✓	✓	✓	✓		✓
Near Stations		✓		✓			✓	
Temperature			✓		✓	✓		✓
Humidity					✓	✓		✓
Wind						✓		
Weekday				✓	✓	✓		✓
Day					✓			✓
Holiday					✓	✓		
Season						✓		
Week						✓		
Redistribution								
Other								
Evaluation	RAE[a]	MAE[b]	MRE [c]	RMSE[d]	RMSE[5.2]	RMSE[5.2]	N/A	VAR[e]

Table 2.1: Variables used in Related Works

[a]Relative Absolute Error
[b]Mean Absolute Error
[c]Mean Relative Error
[d] Root Mean Squared Error
[e]Explained Variance

Chapter 3

Data

This project uses a wide variety of data obtained from sources such as public datasets, information requests to UK authorities and web Application Programming Interfaces (APIs). A description of how this data was collected, stored, cleaned and combined is outlined in this chapter.

3.1 Bicycle Availability Data

In 2015, TfL made available a web API which provides a unified and simple way to access real-time status information of the various modes of transport in the city, including Santander Cycles. Via the `https://api.tfl.gov.uk/bikepoint` endpoint of this API, one can access live bicycle and docking space availability for all docking stations in the sharing scheme. Additional information such as the station's location, installation date and size are also provided. The data, available 24/7 in JavaScript Object Notation (JSON) and Extensible Markup Language (XML), is updated every 5 minutes. It was decided to use the JSON format as it is easier to read and uses less memory than it's markup counterpart. A simplified JSON response from the Bikepoint endpoint can be observed in listing 1.

Since TfL's unified API does not support historical data queries, the live data had to be ingested for it's analysis. For this purpose, we developed a shell script which, when executed, fires a HTTP GET request the Bikepoint API endpoint and stores it's JSON response under a folder in the file system. The script was scheduled to run every 5 minutes in an Amazon Elastic Compute Cloud Linux instance using the cron daemon. Given that the average response size is 1.7 MB, around 490 MB of data were collected daily. It was quickly discovered that some of endpoint's responses were

```
[
  {
    ...
    "id": "BikePoints_1",
    "url": "https://api-argon.tfl.gov.uk/Place/BikePoints_1",
    "commonName": "River Street , Clerkenwell",
    "placeType": "BikePoint",
    "additionalProperties": [
      {
        ...
        "category": "Description",
        "key": "NbEmptyDocks",
        "sourceSystemKey": "BikePoints",
        "value": "15",
        "modified": "2016-08-03T12:35:54.437Z"
      },
      ...
      {
        ...
        "category": "Description",
        "key": "NbDocks",
        "sourceSystemKey": "BikePoints",
        "value": "19",
        "modified": "2016-08-03T12:35:54.437Z"
      }
    ],
    ...
    "lat": 51.529163,
    "lon": -0.10997
  },
  ...
]
```


Listing 1: Simplified JSON response from the Bikepoint endpoint of TfL's unified API.

corrupt or incomplete, so we included functionality to check the response for errors and to retry the request if needed. A second shell script was scheduled to run every 2 hours using this same daemon to, first, group and compress all the responses done in this time period, and second, to upload them to the Amazon Simple Storage Service for it's secure and reliable preservation. Using Amazon's cloud computing services allowed us to have a secure and reliable infrastructure to run our scripts and store the data without the need to manage it's hardware and software. In this way, we collected 6 weeks of data ranging from the 15th of May to the 26th of June, 2016. Although there is no agreed convention for the amount of data needed to accurately predict bicycle availability, as the previously discussed works use sizes ranging from just 3 weeks up

to 2 years, the study done by Froehlich et al [12] concludes that 10 to 15 weekdays of data are enough. Therefore, we believe that our data is significant enough for it's analysis.

All the collected JSON files were then parsed and rearranged into dataframes conforming to the principles of Tidy Data [28], which are that each variable should be represented as a column, each observation as a row and each observational unit as a table. Tidying has the advantage of making the dataset easier to manipulate and visualize while also keeping a low memory footprint. For this, we used Pandas [16], a high-performance library that provides data structures and tools for data analysis in the Python programming language. The structure of the two resulting dataframes, readings and stations, along with their first 5 entries, is outlined in figures 3.1 and 3.2, respectively. Note that the timestamps are provided in the Coordinated Universal Time (UTC) to avoid timezone confusions.

		NbBikes	NbDocks	NbEmptyDocks	NbUnusableDocks
Id	**Timestamp**				
BikePoints_1	**2016-05-16 07:01:29.163000064+00:00**	10.0	19.0	8.0	1.0
	2016-05-16 07:11:30.432999936+00:00	8.0	19.0	10.0	1.0
	2016-05-16 07:16:30.956999936+00:00	7.0	19.0	11.0	1.0
	2016-05-16 07:26:32.369999872+00:00	5.0	19.0	13.0	1.0
	2016-05-16 07:31:33.192999936+00:00	6.0	19.0	12.0	1.0

Figure 3.1: First entries of the "Bicycle Availability Readings" dataframe

	Name	Latitude	Longitude	InstallDate	Installed	Temporary	RemovalDate
Id							
BikePoints_1	River Street, Clerkenwell	51.529163	-0.109970	2010-07-12 15:08:00	True	False	NaT
BikePoints_10	Park Street, Bankside	51.505974	-0.092754	2010-07-04 11:21:00	True	False	NaT
BikePoints_100	Albert Embankment, Vauxhall	51.490435	-0.122806	2010-07-14 09:31:00	True	False	NaT
BikePoints_101	Queen Street 1, Bank	51.511553	-0.092940	2010-07-14 10:18:00	True	False	NaT
BikePoints_102	Jewry Street, Aldgate	51.513406	-0.076793	2010-07-14 10:21:00	True	False	NaT

Figure 3.2: First entries of the "Docking Stations" dataframe

While the developer documentation of TfL's unified API states that a station's bicycle availability readings are updated every 5 minutes, a simple inspection to a random sample of the readings reveals that this is not entirely true (figure 3.1). We suspect then, that a station publishes a status update only when a change in the station occurred in the last 5 minutes. This would mean, first, that most of the readings will be different than the previous one (except for the case in which an equal number of bikes left and

arrived to the station during the 5 minute window, leaving the station in the same state as it was before but triggering a status update), and second, that the count of readings updates of each station reflects how active the station is.

To verify this hypothesis we computed the measure Readings Count (RC) of each station, which is the number of readings updates that occurred during the 6 weeks time period, to get a set of the 100 most active stations according to this metric. We then compared this set to the set conformed by the 100 stations with the highest average Binary Activity Score (BAS) during this same period, which showed that only 5 stations were not shared by these two rankings. While RC and BAS are not directly comparable as BAS ignores the previously discussed scenario when a station ends up with the same number of bicycles even though some activity occurred, we conclude that this result is significant enough to verify our hypothesis.

An analysis of the stations dataframe reveled that some of the stations data was noisy due to inconsistencies, relocations and renamings. At this point, to avoid deleting meaningful data which might be useful later, it was decided to remove only blatant errors such as a station located outside Greater London or stations duplicates that had the same Id but different names, locations, or installation dates, and to perform further ad-hoc cleansing during the analysis and modeling tasks.

The readings dataframe also contained inconsistencies which we suspect were caused due to closures or maintenance works, as our first analysis showed that there were 53 stations that did not reported any readings during one or more days (figure 3.3). Furthermore, it was also observed that a station's status is not updated exactly every 5 minutes, but that the period between readings varies from a couple of milliseconds (in most of the entries) to up to 2 minutes, which causes discrepancies and noise in the moments in which the readings are updated. The former issue was only noted, because as done with the stations dataset, further cleansing was delayed to future phases.

After this first analysis and cleansing, our bicycle availability dataset ended up consisting of 1,500,928 readings of 779 docking stations.

3.2 Bicycle Redistribution Data

To mitigate the imbalances of bicycle distribution in the docking stations (which will be discussed with greater detail in the Analysis chapter), the service operator deploys teams that redistribute the bicycles by collecting them from one station and taking them to another using vans. A key contribution of this project is to study the inclusion of

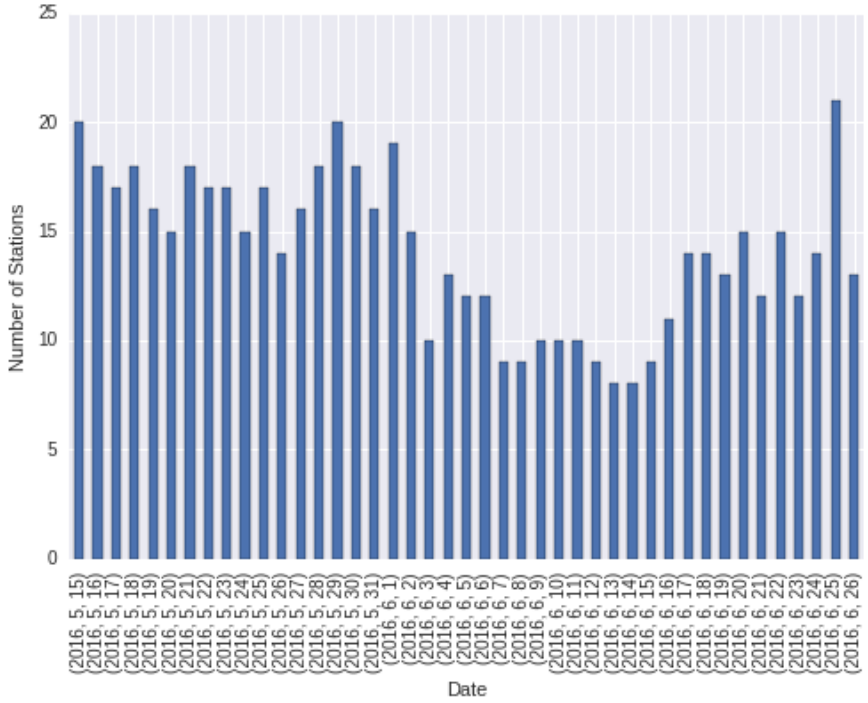

Figure 3.3: Number of docking stations that did not produce any updates for a given day in the 6 weeks time period.

features that model this redistribution activity in the predictors, as, to the best of our knowledge, no other work has done it before. This could be because the data is not publicly available or is usually not logged by the service operators. For example, in the case of Santander Cycles, even though the scheme has been operating since 2010, redistribution data before June 2015 is not available.

This data was obtained by making a Freedom of Information request to TfL through the WhatDoTheyKnow website, which anyone can use to ask for any information held by a public UK authority. Via the request with reference number FOI-2551-1516, TfL provided us data about each collection and distribution event that occurred in the stations of Santander Cycles during the first 6 months of the year 2016, specifically, the number of bicycles collected or dropped, the name of the station and the time of occurrence (UTC). The merging of this data with the bicycle availability dataset was not straightforward due to inconsistencies in the names of the stations, e.g. the former's station "Langdon Park" is called "Langdon Park, Poplar" in the latter. After manually matching these problematic stations and removing the ones that could not be matched, we ended up having 43,396 collection events in 756 stations and 39,677 distribution events in 768 stations. As before, the data was tidied up and rearranged

into two dataframes, one for each type of event.

		NbBikes
Id	Timestamp	
BikePoints_1	2016-01-02 18:19:00+00:00	6
	2016-01-05 10:49:00+00:00	10
	2016-01-06 14:25:00+00:00	7
	2016-01-08 11:14:00+00:00	12
	2016-01-10 11:33:00+00:00	9

Figure 3.4: First 5 entries of the ”distributed” dataframe.

3.3 Weather Data

Given that bicycle usage is heavily influenced by the weather conditions [26], it was decided to study it's effect on bicycle availability throughout the stations of Santander Cycles. London's historical weather data is available for download from the `http://api.wunderground.com/api/KEY/history_DAY` endpoint published by The Weather Underground. Using this endpoint, we obtained, in JSON format, three weather condition readings per each hour (at 0, 20 and 50 minutes) of every day between the 15th of May to the 26th of June, 2016. While the measurements were done by the weather station installed in the Heathrow Airport (located 16 miles away from central London), we assume that they represent the weather conditions in the city fairly well. The JSON files were then parsed and structured into a Pandas dataframe with the following data: temperature (°C), dew point (°C), humidity (%), pressure (mBar), visibility (km), wind speed, wind direction (°), rain (boolean), fog (boolean), tornado (boolean), hail (boolean) and timestamp (UTC).

	DewPt	Fog	Hail	Humidity	Pressure	Rain	Snow	Temp	Thunder	Tornado	Visibility	WindDirD	WindSpeed
Timestamp													
2016-05-14 23:00:00+00:00	4.0	0	0	61.0	1023.0	0	0	9.0	0	0	25.0	50.0	5.6
2016-05-14 23:20:00+00:00	3.0	0	0	71.0	1023.0	0	0	8.0	0	0	10.0	100.0	5.6
2016-05-14 23:50:00+00:00	3.0	0	0	71.0	1023.0	0	0	8.0	0	0	10.0	80.0	5.6
2016-05-15 00:00:00+00:00	4.0	0	0	74.0	1023.0	0	0	7.0	0	0	25.0	90.0	5.6
2016-05-15 00:20:00+00:00	3.0	0	0	71.0	1023.0	0	0	8.0	0	0	7.0	0.0	3.7

Figure 3.5: First 5 entries of the ”weather” dataframe.

A first analysis of the weather dataset revealed that 2 dew point, 1 humidty, 1 temperature, 17 visibility, 1 wind direction and 1 wind speed readings were missing.

Given that these missing values were very sparse and that the immediate previous and next readings were available, it was decided to impute them using linear interpolation. No further data cleansing was needed, which led us to end up having 3,150 entries in this dataset.

3.4 Merged Datasets

The bicycle availability, weather conditions and bicycle redistribution datasets were merged together into a single time series dataset for it's use in the analysis and prediction phases of the project. To fix the irregular intervals found in the raw data, this new merged dataset was resampled using a frequency of 5 minutes which transformed it into a regular time-series dataset, which enabled a more accurate and less noisy modeling by making readings occurring at the same time but in different days directly comparable. This regularity was also needed to correctly assess the stations using the metrics defined in the previous chapter such as NAS and NAB. Note that the assessment of the bicycle sharing scheme's performance using the performance indicators defined in the service level agreement was done using the original irregular time-series dataset to make it as accurately as possible.

Chapter 4

Santander Cycles Analysis

One of the goals of this project is to achieve a deeper understanding about Santander Cycles' to empower transport authorities and service operators to develop data-driven strategies that improve the commuting experience through the city. To support this goal, this chapter presents the patterns, statistics and service levels discovered by analyzing the data described in chapter 3.

4.1 Overall Station Activity

Figure 4.1 uses each station's DayView NAS to display the overall system activity during weekdays and weekends. This metric was chosen over RC as it takes into account the number of bicycles involved in the station's activity as well as it's size. For weekdays, the plot reveals two large spikes corresponding to the morning and evening commute periods, and a third smaller spike during lunchtime. It is no surprise that these spikes are not present during weekends, which instead show higher demand from 12:00 to 16:00, indicating a more casual usage in this period. However, it should be noted that the stations are fairly active during the weekend too. The shaded areas in the figure represent confidence interval bands set at 68%, that is, one standard deviation from the mean. The tightness of the bands indicate that these patterns are shared by most of the stations.

To spatially compare the stations' activity levels, we computed the mean DayView NAS of each station for weekdays and weekends, which we then re-scaled to numbers between zero and one for their representation with a heat map (figure 4.2). In the weekdays map, it can be observed that the stations become more active as they get closer to the city center, with stations near Hyde Park, Convent Garden, Westminster, Waterloo

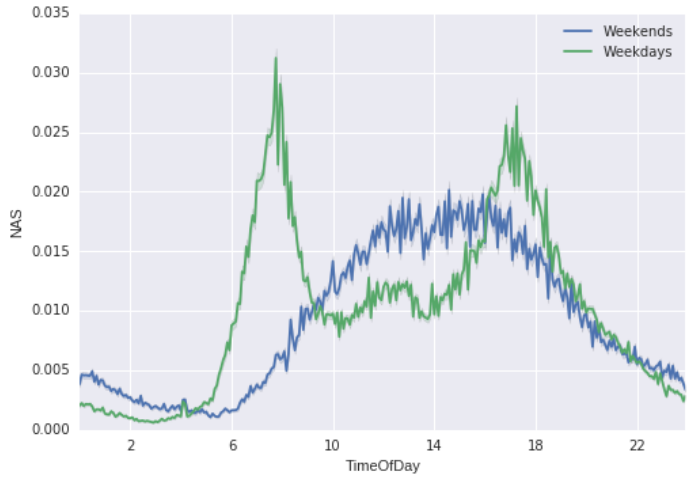

Figure 4.1: Overall system's activity (as measured by NAS) during weekdays and weekends.

Figure 4.2: Average station activity during weekdays (left) and weekends (right). Red indicates higher activity levels.

Station, Liverpool Street and Barbican Station, having particularly high activity levels. This is most probably caused by the commute to work pattern and business activity being concentrated in the city center. During weekends, the activity spikes around places for leisure, such as Hyde Park and Trafalgar Square. Not surprisingly, stations near business areas are not as active as non-work days.

4.2 Overall Bicycle Availability

Inevitably, commuting pattern will cause imbalances in bicycle distribution. This means that some stations will become fuller than others at different times of the day depending on their location. Given that the amount of bicycles available in the sharing scheme is fairly stable from day to day, this imbalances are the main culprits for negative usage experience, mostly caused when a user tries to take or return a bicycle and

Figure 4.3: Bicycle availability at 7:00 (top left), 10:00 (top right), 16:00 (bottom left) and 19:00 (bottom right) of Monday 16 March 2016. Red colors indicate fuller stations.

finds the station empty or full, respectively. To observe this spatio-temporal distribution patterns, we have produced plots which show the stations' NAB (a measure of it's fullness) before and after the morning and evening rush hours during weekdays and in the morning and evening of weekends (figure 4.3).

Figure 4.3 illustrates that, at the beginning of the morning peak period, bicycles are mostly located in the surroundings of the city centre, where stations of particularly high availability can be found in Camden, Earl's Court, Nine Elms, North Kensington, Elephant and Castle and Westbourne Green, while at the end of the morning peak time period, bicycles concentrate heavily in the city centre, specially in the Knightsbridge, Westminster, Mayfair, St Jame's, Piccadilly Circus, Fitzrovia, Convent Garden, Temple, Clerkenwell, Southbank and Bank areas. Between the end of the morning peak time period and the beginning of the peak time period, availability in the city centre spreads more evenly, something we suspect is caused by casual trips done by tourists or other non-member users. Later, during the evening peak period, bicycles are taken outside the city centre, with stations located near other means of public transport such as Waterloo Station, Elephant and Castle, Kings Cross Station being particularly full. The observed behavior bolsters the commute pattern hypothesis and suggests that Santander Cycles is also heavily used as a last mile transport option, helping it's users get

from a public transportation hub to their final destination and vice versa.

4.3 Individual Station Availability

Bicycle availability patterns vary widely from station to station, being mostly influenced by the station's location. Figures 4.4 and 4.5 show daytime bicycle availability from May 16 to May 23, 2016, for stations Bank Of England Museum, Bank and Curlew Street, Shad Thames, respectively. In them it can be observed that during weekdays, the former station, which is located in the city centre, maintains a fairly high number of bicycles from 7:00 to 17:00, followed by a sharp drop that makes availability plummet to zero around 17:20. Later, the number of bicycles in the station spikes slightly approximately at 19:30 to then become mostly zero again. This contrasts with the pattern exhibited by the latter station during weekdays, which shows two large spikes at 7:00 and 19:00, but low availability in-between. In both cases, approximately the same pattern occurs during all working days except Wednesday 18 May, which shows atypical behavior. It is difficult to know exactly what the cause of this anomalous behavior could be, as this day is not a bank holiday, had no particularly bad weather and, to the best of our knowledge, saw no major event occurring in London. This atypical behavior is unique to this week and is presented here as proof of the unknown factors that influence bicycle availability in the stations of Santander Cycles, which create significant challenges when modeling their behavior.

4.4 Bicycle Redistribution

To prevent users from running into the undesirable situations of finding no available bicycles or free docking spaces in a station, the service operator redistributes bicycles among stations to keep them from being empty or full for large periods of time. The redistribution is carried out using 34 vans which move approximately 2,170 bicycles per day. As it can be observed in the heavily right skewed histograms of the total number of bicycles collected and distributed in each station shown in figure 4.6, a small number of stations concentrate most redistribution activities. A log transformation was then applied to visualize this information in a heat map (figure 4.7). The map shows that stations from which the most bicycles were collected from are located in the perimeter of the city centre, while the stations with most bicycles dropped to do not follow any evident geographical pattern. Note that, apart from these highly redistributed stations,

Figure 4.4: Daytime bicycle availability in station Bank Of England Museum, Bank from May 16 to May 23, 2016. Note that this station is located in the city centre.

the redistribution activities are spread fairly evenly throughout the remaining stations.

Appendix figures A.1 and A.2 show the time of occurrence of bicycle redistribution events in the top 12 stations that concentrate most redistribution activities. In them we can observe that, while most redistribution events in each station follow a regular pattern, there are some that do not, which we believe has the following consequences when modeling the stations' behavior: First, that the models will learn the increase or decrease in the number of bicycles at the moments where redistribution happens regularly without the need of features representing these activities (as they will be indirectly reflected in the number of available bicycles at those times). And second, that it will be in the moments when abnormal redistribution activities occur that the models will be benefited from the use of features which model these activities.

4.5 Key Performance Indicators

The Service Level Agreement (SLA) [9] signed by TfL and Serco (the service operator) outlines 34 performance indicators (PI) designed to measure the quality of the Santander Cycles' service. Given that bicycle distribution plays such an important role in the overall customer experience, in this project we focus on the PIs that are related to it, PIs 24, 25, 26 and 27. These PIs make the following distinctions of time and

Figure 4.5: Daytime bicycle availability in station Curlew Street, Shad Thames from May 16 to May 23, 2016. Note that this station is located in Bermondsey, outside the city centre.

station priority:

- **Station Priority:** Each docking station was classified by TfL as Priority 1 (P1) or Priority 2 (P2) based on it's importance for the overall operation of the bicycle sharing scheme. Highly used docking stations were classified as P1, while the remaining stations as P2. The maximum number of stations classified as priority 1 should not exceed 100. Note that while the station's priority data is not available through TfL's unified API, a Freedom of Information request answered in November 2015, provided the stations' classification used in this analysis.

- **Peak Hours:** The hours during which crowding on the public transportation system and the traffic on roads are the highest are called Peak Hours. The SLA defines them to be from 7:00 to 10:00 and from 16:00 to 19:00. All other times are considered non-peak hours. It is not specified if peak hours apply during working days only.

PIs 24 and 25 measure the combined number of minutes during which Santander Cycles' stations are empty or full during peak and non-peak hours over a calendar day. In order to do so, the number of minutes for which each docking station is full or empty is recorded, split, grouped and summed by calendar day, priority and peak

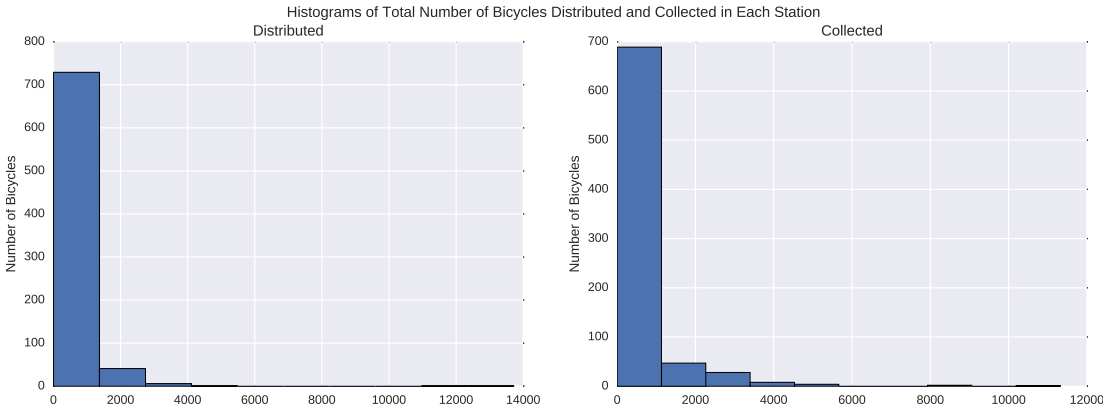

Figure 4.6: Geographical representation of the total number of bicycles distributed (left) and collected (right) from Santander Cycles' stations. The colors range from blue (low number of bicycles) to red (large number of bicycles).

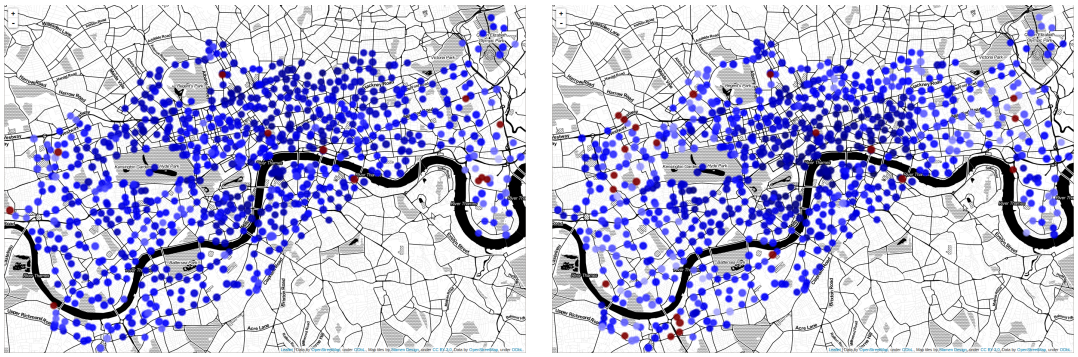

Figure 4.7: Heat map representation of the total number of bicycles distributed (left) and collected (right) in each docking station during 2016 (up to the 26th June). The colors range from blue (low number of bicycles) to red (large number of bicycles).

and non-peak hours to assess the goodness of the bicycle distribution according to the thresholds defined in table 4.1. PIs 26 and 27 measure the number of times that a P1 station is empty or full for more than 30 continuous minutes during peak hours. We use these 4 PIs to assess the performance of the bicycle hire system from 16 May to 26 June, 2016.

4.5.1 Data Cleansing

The data showed that several stations were empty or full during entire days. It is very unlikely that these stations did not have any activity during these time periods as the stations had fairly regular activity before and after. Therefore, we assume that these empty periods were caused by malfunctions in the stations or in the data collection

Priority	Time	Acceptable Service Level
1	Peak Hours	Less than 1,000 minutes for all stations per peak period
1	Non-Peak Hours	Less than 3,000 minutes for all stations
2	Peak Hours	Less than 9,000 minutes for all stations per peak period
2	Non-Peak Hours	Less than 18,000 minutes for all stations

Table 4.1: Number of accumulated minutes needed to meet the acceptable service levels defined by PIs 24 and 25. Note: these acceptable service levels were taken from the original SLA published in 2009 when there were only 352 docking stations.

process and thus should be discarded. To remove this erroneous readings, we deleted any empty or full periods of more than or equal to 720 minutes (12 hours). This number is what we consider the worst case scenario for a naturally inactive station. It was computed by taking into account the 10 hours restriction placed by some boroughs that forbids the service provider to redistribute bicycles between 22:00 and 8:00, and the 2 hours that might be needed for the redistribution to take place in this areas.

4.5.2 Accumulated Empty and Full Minutes

Let us first examine how stations are distributed throughout the city based on the accumulated number of minutes they were empty and full during the 6 week observational period. As it can be observed in figure 4.8, stations that accumulated the most empty minutes are located in the city centre, while stations with more full minutes are located in it's perimeter, particularly in the south. A visual representation of the empty and full periods of the top 20 stations that accumulated the most empty and full minutes during each of the first 6 observed days can be found in appendix figures A.3 and A.4, respectively. These figures suggest that preventing the stations from becoming empty or full after the evening peak hours is particularly problematic for the service operator. Listings of the top 20 most problematic stations for both scenarios (**??** and **??**) can also be found in the appendix

4.5.3 Performance Indicators 24 and 25: Daily Accumulated Empty and Full Minutes

Here we examine if the service levels defined by PIs 24 and 25 (as specified in table 4.1) were met for each day of the observed 6 weeks. To account for the increase in

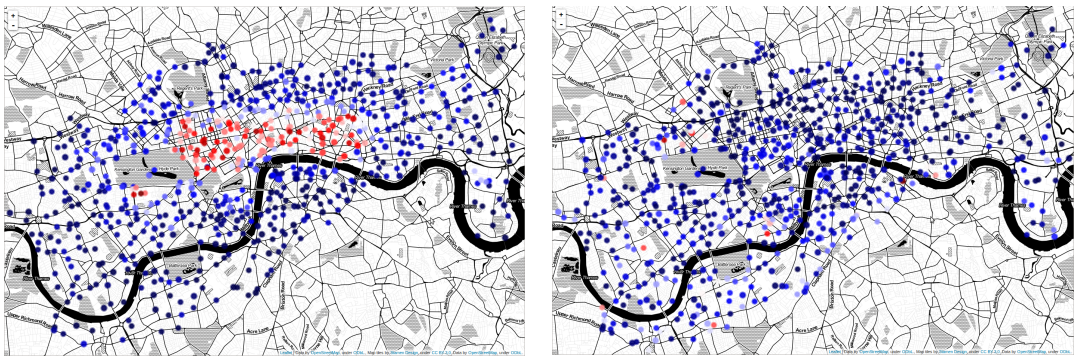

Figure 4.8: Representation of the accumulated number of minutes that the stations were empty (left) or full (right) during the observational period. The colors range from blue to red, where red colors indicate more accumulated minutes.

the number of stations since the P2 thresholds were first set in the SLA, we derived new thresholds for the peak and non-peak hours of these stations (20,000 and 40,000, respectively) using equation 4.1 and then, truncating the resulting number to the nearest thousand.

$$\text{P2Threshold}_{\text{new}} = \frac{\text{P2Threshold}_{\text{old}} * \text{NumberOfStations}_{\text{new}}}{\text{NumberOfStations}_{\text{old}}} \tag{4.1}$$

Table 4.2 summarizes the failure rates for the expected service levels defined by these PIs. In it, we can observe that keeping stations from becoming empty during non-peak hours is particularly problematic for the service operator, with the expected operational levels not being met 100 % and 85.71 % of the times for P1 and P2 stations, respectively. While the system performs better in the scenario of full stations during non-peak hours, it still fails 42.85 % and 30.95 % of the observed days. Furthermore, there is a notorious difference in the system's performance for P1 stations during morning and evening peak time periods, with the service being better during the latter. We suspect that this could be caused by the restriction placed by some boroughs to not perform bicycle redistribution activities at night, which makes it difficult to prepare for the next morning rush hour. Notably, the service levels for full P1 stations during morning and evening peak hours is particularly high, as it being met 86.72 % and 100 % of the times, respectively. Finally, given that P2 stations have less activity than P1 stations during peak hours, it is not surprising that the service operator handles them much better. Here, it can be seen that the acceptable level for full stations in both peak periods is met 100 % of the time, and that, for empty stations, it is also constantly being met, failing only 19.04 % and 4.76 % of the observed days during morning and

Period	Priority	Status	Failure Rate
Non-Peak	1	Empty	100 %
Non-Peak	1	Full	42.85 %
Non-Peak	2	Empty	85.71 %
Non-Peak	2	Full	30.95 %
Morning Peak	1	Empty	76.19 %
Morning Peak	1	Full	14.28 %
Evening Peak	1	Empty	50 %
Evening Peak	1	Full	0 %
Morning Peak	2	Empty	19.04 %
Morning Peak	2	Full	0 %
Evening Peak	2	Empty	4.96 %
Evening Peak	2	Full	0 %

Table 4.2: Failure rates for PIs 24 and 25 in their different scenarios.

evening peak time hours, respectively.

Appendix figures A.5, A.6 and A.7 graphically illustrate the accumulated number of minutes during which P1 and P2 stations were empty or full during non-peak hours, P1 stations were empty or full during peak hours and P2 stations were empty or full during peak hours, respectively.

4.5.4 Performance Indicators 26 and 27: Station Full or Empty Maximum Time Period

PIs 26 and 27 measure bicycle distribution by counting the number of times that a P1 station is full or empty for more than 30 consecutive minutes (lower counts indicate better performance). Figure 4.9 shows the number of empty and full continuous periods that exceeded the maximum allowed limit during each observed day. In both cases we can observe that there are more threshold violations during the morning peak hours than during the evening peak hours. There is no other particular pattern about when the violations occur, which makes planning for redistribution activities difficult.

The mean number of violations per station is 12.34 (s = 9.78) for the empty station scenario and 7.25 (s = 7.19) for the full station scenario. Such high standard deviations mean that not all stations violate the threshold with similar frequency. A further

Figure 4.9: Number of periods exceeding the maximum allowed number of minutes for empty (top) and full (bottom) stations during each day.

inspection of the quartiles (figure), shows that a relatively low number of stations is causing the large majority of the violations, which means that focusing the redistribution efforts on these problematic stations could achieve a significant increase in the system's performance in PIs 26 and 27. Performance could be further increased by focusing on the evening peak time periods, as this is when most violations occur. Appendix figures A.8 and A.9 show the number of times that each station violated the empty and full thresholds, respectively.

The maps in figure 4.11 show if a station was more frequently violating PIs 26 and 27 during the morning or evening peak hours for both empty and full station scenarios. It can be observed that most of the stations that were empty during morning peak hours are mostly located outside the city centre or near public transportation hubs such as Waterloo Station, while empty stations during evening peak hours are mostly located in the city center. The map for the maximum full station period scenario shows the inverse behavior, which showcases again the commute pattern in Santander Cycles.

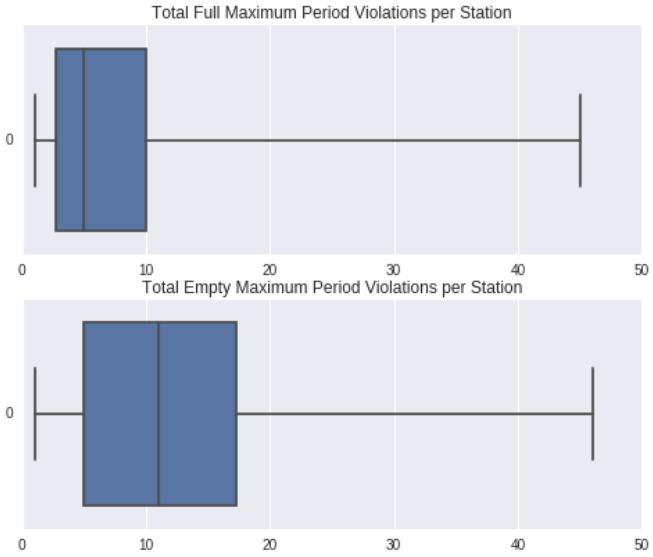

Figure 4.10: Box plots of the number of violations of the maximum full (top) and empty (bottom) periods per station.

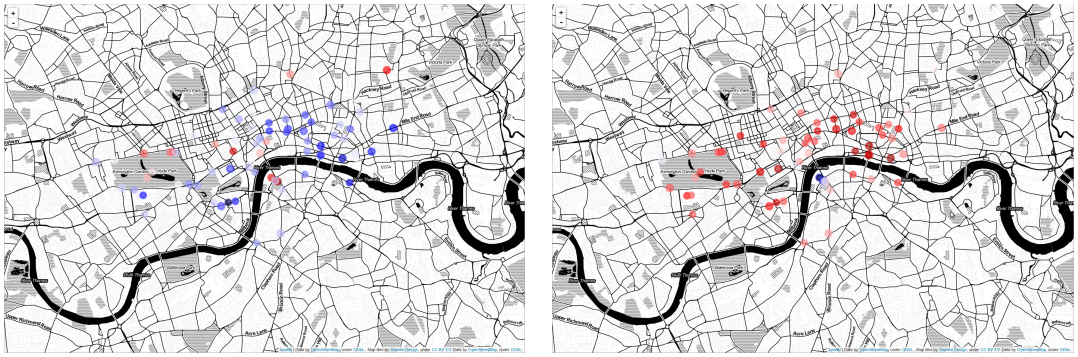

Figure 4.11: Representation of the stations violating the max allowed empty (left) or full (right) number of continuous minutes during peak hours. Red colors indicate stations that are more frequently empty during or full morning peak hours, while blue colors during evening peak hours

.

Chapter 5

Methodology

5.1 Hypothesis

Machine learning models can be used to predict bicycle availability in the docking stations of London's bicycle sharing scheme, Santander Cycles. The inclusion of exogenous features such as weather conditions, bicycle redistribution activity, surrounding stations behavior and type of day (holiday, weekend, or weekday) as input to the predictor can further improve it's performance. Similarly, by preventing overfitting, the use of regularization techniques can also help the model generalize better to unseen data. Accurate forecasts would be greatly valuable to both users and service operators, allowing the former to plan their journey better to prevent running into a full or empty station, and the latter to devise data-driven strategies which encourage bicycle usage, improve customer satisfaction and reduce customer churn.

5.2 Problem Description

The problem of predicting the number of available bicycles in a station (or inversely, the number of free docking spaces) n minutes ahead, can be approached as a regression problem, in which the predictor takes in various input features, such as weather, previous availability numbers, redistribution activity and surrounding stations readings, to output the expected number of available bicycles. In other words, regression models how changes in the independent variables (input features) affect the value of the dependent variable (number of available bicycles) [19], which can be mathematically stated as follows:

Let nb^s_{t+n} be the number of bicycles available n minutes ahead from time t in station

s. Then, we aim to find a function f that relates the independent variables X, unknown parameters β and noise η to nb^s_{t+n} adequately.

$$nb^{id}_{t+n} = f(X, \beta, \eta) \tag{5.1}$$

In this work, we experiment with various functions and input features to find the combination that produces the most accurate results.

5.3 Evaluation

The performance of our various predictors is measured using Average Root Mean Squared Error (RMSE_{AVG}), which is computed by averaging the Root Mean Squared Error of the predictor for every docking station:

$$\text{RMSE}_{\text{s}} = \sqrt{\frac{1}{I}\sum_{i=1}^{K}(t_i - y_i)^2} \tag{5.2}$$

$$\text{RMSE}_{\text{AVG}} = \frac{1}{S}\sum_{s=1}^{S}\text{RMSE}_{\text{s}} \tag{5.3}$$

where y_i is the number of available bicycles predicted by the model for observation i, t_i the real number of available bicycles for observation i, I the number of observations for station s and S the set of stations to predict availability for. RMSE, which has been frequently used as the evaluation metric in related works such as [32], [5] and [13], has the property of harshly penalizing large deviations between the observed and the predicted values, and the advantage of reporting the error in the same units as the predicted value (number of bicycles), something that makes it particularly suitable for our business scenario as it is easy to explain and understand.

Furthermore, given that the size of the docking stations in Santander Cycles varies widely (ranging from 6 to 123 docking spaces), we use an additional metric that takes station size into account: Normalized Average Root Mean Squared Error ($\text{NRMSE}_{\text{AVG}}$). This metric expresses the error in terms of the percentage of the station size by which the predictor was wrong and is computed by averaging the result of dividing each station's RMSE_s by it's size Size_s:

$$\text{NRMSE}_{\text{AVG}} = \frac{1}{S}\sum_{s=1}^{S}\frac{\text{RMSE}_{\text{s}}}{\text{Size}_s} \tag{5.4}$$

5.4 Data

To train and evaluate our models, the six weeks of ingested data are split as follows (as suggested by [31]):

1. **Training Set**: Readings collected between the 16th of May and the 12th of June. This data is fed to the model during the training phase so that it learns it's patterns.

2. **Validation Set**: Data observed from the 13th to the 19th of June 2016. It is used to tune the model's parameters to find the optimal values and prevent it from overfitting.

3. **Test Set**: Readings collected between the 20th and the 26th of June 2016. We use this data to establish the performance of the predictor in unseen data.

Dividing the data as just described allows the fine-tuning of the models' hyperparameters in a way that avoids patterns of the unseen data to leak into the training of the models (which would make the evaluation metrics no longer trustworthy) and also the accurate assessment of their generalization performance.

5.5 Predictors

We experiment with the predictors outlined in this section to solve the problem of predicting bicycle availability throughout the docking stations of Santander Cycles.

5.5.1 Baseline Predictor

Simple models that follow straightforward logical rules based on the current and historical readings of a docking station are used to establish the minimum performance that can be expected from other more complex predictors. Here, we use the following baseline model:

- **Historical Average Model (Hist AVG)**. Predicts that the number of available bicycles at time t is the mean of the observed values at time t in previous days.

5.5.2 Linear Regression Model

Linear Regression (LR) models the relationship between a dependent variable y and a set of N independent variables $(x_1, x_2, ..., x_n)$ as

$$y = \beta_o + \sum_{n=1}^{N} x_j \beta_i \tag{5.5}$$

where β_i are parameters or coefficients that determine the influence of each independent variable on the value of the dependent variable [11]. β_0 is called the intercept of the model, it determines the mean output value when all the independent values are zero and helps to account the effect of any independent variables omitted by the model. Here, these coefficients are estimated from the data using the least squares approach, which finds the set of coefficients that minimizes the residual sum of squares (RSS) as shown in equation 5.6.

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{K} (t_i - y_i)^2 \tag{5.6}$$

As it's name suggests, this model makes the assumption that the relationship between the variables is linear, or that this assumption provides an acceptable approximation. This means that if x_i changes by Δx_i, the expected value of y changes by a proportional amount determined by $\beta_i \delta x_i$. Other assumptions made by the LR model are that the expected response value is explained by the sum of the separate effects of the independent variables (additivity) and that that the deviations between the perdicted and the observed value are not autocorrelated, have homoscedasticity (have equal variance) and are normally distributed.

Despite the limitations imposed by it's strong assumptions, we use Linear Regression as it has the advantages of being simple, fast and interpretable, which make it the perfect choice if it achieves an acceptable performance. Furthermore, it is a primitive type of our other more complex model, which will help assess the influence of non-linearities in the feature space and understand the fundamental differences between them.

5.5.3 Generalized Additive Model

Generalized Additive Models (GAM) are a type of likelihood-based regression models developed by Trevor Hastie and Robert Tibshirani [14] that relate the value of a dependent variable y to the sum of smooth functions of the independent variables x_i. GAMs

have the following general form

$$g(\mathrm{E}(y)) = \sum_{i=1} s_i(x_i) + \varepsilon \tag{5.7}$$

where s_i are the smooth functions, ε independent random noise and g a link function that links the expected value of the dependent variable y to the predictor variables. They have the advantages of being

1. **Flexible:** GAMs can capture complex nonlinear patterns by modeling them with smooth functions.

2. **Automatable:** During the training phase, the smooth functions are automatically derived from the data, which has the advantage of not having to know upfront the underlying patterns between the variables.

3. **Interpretable:** Since in additive models the marginal impact of a variable is not influenced by the values of other variables, one can easily observe the contribution of each variable to the response.

4. **Regularizable:** The smoothness of the functions can be controlled with parameters to prevent the model from overfitting the data.

which make them a well balanced model that stands in the middle ground between extremely flexible black box models such as Neural Networks and rigid interpretable models like Linear Regression.

The smooth functions of the input predictors can be represented using splines or tensor products of splines (when modeling the joint effect of two or more variables) [29]. Splines, which are curves made up of sections of polynomials of degree n joined together in points known as knots, are used to describe the patterns exhibited by the input variables so that the combination of them explains the response variable. They can be classified as smoothing and regression based, where the former puts knots at every data point (which is wasteful as, due to the penalization term, the resulting spline will always be smoother than what the degrees of freedom specified by the knots suggest), while for the latter, the knots' location must be chosen (typically evenly spacing them through the range of values or placed at quantiles). Here we will use penalized regression splines as they are cheaper to compute and can be represented as the linear combination of a set of basis functions (determined by the number of inner knots and the order of the spline), which has the added advantage of allowing direct interpretation of the parameters. The smoothness of the spline functions is usually determined

by the number of knots, where fewer knots mean more smoothness. However, here we use a type of regression splines known as penalized regression splines, which add a term λ to the least squares minimization objective to penalize wiggliness, where $\lambda = \infty$ results in maximum smoothness (a straight line) and $\lambda = 0$ in an unpenalized spline.

The following equation was proposed by Chen et al [5] to predict bicycle availability using a GAM;

$$\begin{aligned} y_t = {} & s_1(x_t^{TimeOfDay}) \cdot 1(x_t^{DayType} = \text{"}weekday\text{"}) \\ & + s_2(x_t^{TimeOfDay}) \cdot 1(x_t^{DayType} = \text{"}weekend\text{"}) \\ & + s_3(x_t^{TimeOfYear}) + \beta_1 \cdot 1(x_t^{Weather} = \text{"}rainy\text{"}) \\ & + \beta_2 \cdot 1(x_t^{Weather} = \text{"}foggy\text{"}) + s_4(x_t^{Temperature}, x_t^{Humidity}) \\ & + s_5(y_{t-1}) + s_6(y_{t-2}). \end{aligned} \tag{5.8}$$

here, y_t is the number of bicycles in a docking station at time t, $s_1, ..., s_6$ are smooth functions of the variables time of day, time of year, weather, temperature, number of bicycles at time $t-1$ and number of bicycles at time $t-2$ and β_1 and β_2 coefficients that control how weather events affect availability. We use equation 5.8 as a starting point for our experiments with this model which then we modify to include new features or to remove not helpful ones.

5.6 Predictions

This section outlines how the previously described models are used to predict bicycle availability in the docking stations of Santander Cycles 10, 60 and 180 minutes ahead. First, we present the available feature space, then, the mathematical relationships between the features and finally, how these two are used in each prediction scenario.

5.6.1 Data Cleansing

For the training and evaluation of our models we removed stations that have less than 29 days of data (69 %).

5.6.2 Features

The data used in these experiments consists of 9,280,524 records of bicycle availability readings for 765 stations collected between the 15th of May and the 26th of June,

2016, and augmented with derived and external variables such as historical average readings, weather conditions and redistribution activities, forming the following **133 dimensional** feature space:

1. **Station Status.** Variables that describe the station's state at time t.
   (a) NbDocks: Total number of docking spaces in the station.
   (b) NbBikes: Number of fully functioning bicycles that are available for users to take.
   (c) NbEmptyDocks: Number of fully functional docking spaces available in which users can leave a bicycle.
   (d) NbUnusableDocks: Number of docking spaces which are not functional.
   (e) HistAvg: Average number of fully functional bicycles available for hire in a station at time t in past days.
   (f) TMinus{k}: Number of bicycles available at the docking station k periods of 5 minutes before, where $K = \{2,3,12,18\}$. For example, TMinus12 is the number of bicycles available 60 minutes before.
2. **Temporal.** Variables that reflect in various ways the temporal moment t in which the readings occurred.
   (a) Timestamp: Date and time of the readings in the UTC timezone.
   (b) TimeOfDay: Number of seconds elapsed since the beginning of the day.
   (c) TimeOfYear: Number of seconds elapsed since the beginning of the year.
   (d) DayOfWeek: Day of the week represented as an integer number (Monday=0 and Sunday=6).
   (e) WeekOfYear: Number of weeks ellapsed since the beginning of the year.
   (f) Holiday: Boolean value representing if the day is a holiday.
   (g) Weekend: Boolean value representing if the day is a weekend.
   (h) Weekday: Boolean value representing if the day is a weekday.
3. **Weather.** Variables that reflect the weather conditions in the city of London as measured by the weather station located in Heathrow Airport.
   (a) Temp: Indicates in °C how hot or cold the day is.

(b) DewPt:Reflects the atmospheric moisture in °C. A value greater than 20 °is considered uncomfortable.

(c) Humidity: Shows the amount of vapor in the air. It gives a sense of the probability of fog, dew, or precipitation occurring.

(d) Pressure: Weight of the air above us, measured in millibars.

(e) Visibility: Reflects the distance in kilometers from which an object can be clearly seen.

(f) WindSpeed: Indicates the speed at which the winds blows in kilometers per hour.

(g) WindDirD: Indicates the direction of the wind in azimuth degrees.

(h) Rain: Boolean value indicating the presence of rain.

(i) Fog: Boolean value indicating the presence of fog.

(j) Tornado: Boolean value indicating the presence of a tornado.

(k) Hail: Boolean value indicating the presence of hail.

4. **Bicycle Redistribution.** Variables that account the redistribution activities done by the service provider to maintain good bicycle distribution throughout the docking stations.

   (a) CollNbBikes: Number of bicycles collected from the docking station.

   (b) CollNbBikesCum{k}: Accumulated number of bicycles collected from the docking station in the past k periods of 5 minutes before, where $K = \{2,6\}$. For example, CollNbBikesCum6 represents the total number of bicycles collected from the docking station in the past 30 minutes.

   (c) DistNbBikes: Number of bicycles distributed to the docking station.

   (d) CollNbBikesCum{k}: Accumulated number of bicycles distributed to the docking station in the past k periods of 5 minutes before, where $K = \{2,6\}$. For example, DistNbBikesCum2 represents the total number of bicycles distributed to the docking station in the past 10 minutes.

5. **Surrounding Stations**

   (a) Near{J}: Number of bicycles available at time t at the j nearest docking station (as ranked by Vincenty distance), where $J = \{1,2,3,\ldots,19,20\}$.

For example, Near3 represents the number of bicycles available in the third nearest docking station.

(b) Near{j}TMinus{k}: Number of bicycles available at the j nearest docking station (as ranked by Vincenty distance) k periods of 5 minutes before, where $J = \{1,2,3,\ldots,9,10\}$ and $K = \{1,2,3,12,18\}$. For example, Near2TMinus2 represents the number of bicycles available in the second nearest docking station 10 minutes ago.

5.6.3 Equations

Using this feature space, we explain the relationship between the number of bicycles available in a docking station y_t (NbBikes) and the input variables x_t at time t with equation

$$\begin{aligned} y_t = {} & s_1(x_t^{TimeOfDay}) \cdot \mathbb{1}(x_t^{Weekday} = 1) + s_2(x_t^{TimeOfDay}) \cdot \mathbb{1}(x_t^{Weekend} = 0) \\ & + s_3(x_t^{TimeOfDay}) \cdot \mathbb{1}(x_t^{Holiday} = 1) + \beta_1 \cdot \mathbb{1}(x_{t-1}^{Rain} = 1) + \beta_2 \cdot \mathbb{1}(x_{t-1}^{Fog} = 1) \\ & + s_4(x_{t-1}^{Temperature}, x_{t-1}^{Humidity}) + s_5(x_{t-1}^{NbBikes}) + s_6(x_{t-2}^{NbBikes}) + s_7(x_t^{DistNbBikes}) \\ & + s_8(x_t^{CollNbBikes}) + s_9(x_{t-1}^{Near1NbBikes}) + \cdots + s_{9+n}(x_{t-1}^{NearNNbBikes}) \end{aligned} \tag{5.9}$$

where s_i and β_i are smooth functions and coefficients of the input variables, respectively, t is the time for which we want to predict bicycle availability, $t-1$ is the hypothetical present time (the moment when the prediction is made) and $t-2$ is some time before $t-1$. The indicator function $\mathbb{1}(condition)$, which evaluates to 1 if it's condition is true and to 0 otherwise, is used to represent that the influence of the time of day is different for holidays, weekdays and weekends, while coefficients β_1 and β_2 account differences in bicycle availability caused by weather events.

Similarly, we use equation 5.10 to denote the same relationship for the LR model

$$\begin{aligned} y_t = {} & \beta_1 \cdot x_t^{TimeOfDay} + \beta_2 \cdot x_t^{Weekday} + \beta_3 \cdot x_t^{Holiday} + \beta_4 \cdot x_{t-1}^{Rain} + \beta_5 \cdot x_{t-1}^{Fog} \\ & + \beta_6 \cdot x_{t-1}^{Temperature} + \beta_7 \cdot x_{t-1}^{Humidity} + \beta_8 \cdot x_{t-1}^{NbBikes} + \beta_9 \cdot x_{t-1}^{NbBikes} \\ & + \beta_{10} \cdot x_t^{DistNbBikes} + \beta_{11} \cdot x_t^{CollNbBikes} + \beta_{12} \cdot x_{t-1}^{Near1NbBikes} + \cdots \\ & + \beta_{12+n} \cdot x_{t-1}^{NearNNbBikes} + \beta_0 \end{aligned} \tag{5.10}$$

where β_0 is the intercept and the remaining β_i are coefficients that measure how the dependent variable y_t is influenced by each input variable.

5.6.4 Implementation Details

We predict bicycle availability in the docking stations of Santander Cycles by fitting a GAM and a LR for each station. Due to a GAM implementation not being available in our programming language of choice (Python), we wrap R's MGCV [29] implementation using a skeleton provided by Python's Scikit-Learn machine learning library [23].

5.6.5 Prediction Scenarios

The time windows presented here have advantages for both users and service operators. Short-term predictions allow the former to plan their upcoming journey better by giving them enough time to walk to a station that has available bicycles (considering a human adult walks 500 meters in 7 minutes [1] and that TfL's density rule states that there should be a station at least every 500 meters), while medium-term and long-term predictions allow the service operator enough time to plan and implement bicycle redistribution activities even in heavily congested areas of the city.

1. **10 Minutes Ahead**

   For short-term predictions we use the full equations described in subsection 5.6.3, where $t-1$ and $t-2$ represent the time 10 and 15 minutes before the predicted time t [5]. Since we assume that the weather conditions do not change significantly in such small time window, the weather variables take the values known at the hypothetical present time $t-1$. Furthermore, we include variables that reflect bicycle availability in the docking station, as well as in the nearest docking stations, at times $t-1$ and $t-2$. Variables that account for redistribution activity happening at the station between times t and $t-1$ are also included.

2. **60 Minutes Ahead**

   For medium term predictions, we use equations 5.9 and 5.10 as in the 10 minutes ahead prediction scenario but without the weather events variables x^{Fog} and x^{Rain} [5]. Given that temperature and humidity can be considered relatively stable within periods of one hour, these variables take the known values at the hypothetical present time $t-1$. In this prediction scenario, $t-1$ and $t-2$ represent the time 60 and 90 minutes before the predicted time t.

3. **180 Minutes Ahead**

For long term predictions, variables representing weather conditions and present and previous availability readings in the target and surrounding stations are no longer included as they can change significantly in such large window of time. However, we include a variable which represents the historical average number of bicycles available at time t in the docking station [5].

For each of the time windows, we use the wrapper method to find the subset of features that provides the best performance. This method, which consists of using the models to rank subsets of features, is a simple and effective way to do feature selection. However, it has the disadvantage of being computationally expensive and therefore, due to time and computational constraints, we perform our experiments with a sample of 100 stations.

5.6.6 Regularization

Overfitting is the undesirable condition when a predictor is so flexible that it learns the patterns and noise specific to the training data rather than the ones of the overall population. This condition makes it perform extremely well when evaluated using the training data but very badly with unseen data (it has a high generalization error). GAMs are prone to overfitting due to the high degree of flexibility and complexity that they can achieve through splines [29]. To avoid running into this issue, which can greatly affect the performance of state of the art predictors in the bicycle availability problem [13], we fine-tune the λ parameter of each of the splines by searching the parameter space using SkLearn's GridSearchCV [23] to find the combination of values that achieves the best generalization performance as determined by 2-fold cross validation. Due to the large range of values that the smoothness parameter λ can take (0 to inf), we first perform a broad search to find the approximate magnitude of each feature's λ, e.g. if it is near 0.2, 0.6, 1, 5 or 10, and then do a narrower search around these values. Due to the expensiveness of searching the smooth parameter space, we perform the search in a dataset consisting of only 1 station and then evaluate the outcome using a sample of 100 stations.

We do not experiment with different types of splines (the MGCV package supports thin plate regression splines, Duchon splines, cubic regression splines, splines on the sphere, P-splines, Gaussian process smooths and Soap film smooths) as, according to Wood [29], it is unlikely that the model's performance depends significantly on this parameter.

Chapter 6

Results and Conclusions

6.1 Results

Here we present the results of the experiments outlined in chapter 5. First, we look at how the feature selection and regularization experiments lead us to find the best performing models for each prediction scenario, and then, we evaluate and analyze the performance of the trained models when predicting bicycle availability in all docking stations of Santander Cycles.

6.1.1 Feature Selection Experiments

In our experiments, we vary the features defined by equations 5.9 and 5.10 to find the subset of features that achieve the best performance for each model in each prediction scenarios (wrapper method). The subsets of features used in the experiments are outlined in the following listing:

1. **Baseline (BASE).** As suggested by Chen et al [5], it consists of temporal features (time of day, holiday, weekend and weekdays), weather condition features (rain, fog, temperature and humidity) and previous availability features. Note that, as described in chapter 5, each prediction scenario has a different baseline.

2. **Surrounding Stations (SURR).** Conformed by features defined in the baseline plus features of previous bicycle availability in the 1st (NEAR1), 2nd (NEAR2), 5th (NEAR5) and 10th (NEAR10) nearest stations.

3. **Bicycle Redistribution (RED).** Features defined in the baseline plus features describing the number of bicycles distributed and collected from the docking

station in the past 5 (NOW), 10 (CUM2) and 30 (CUM6) minutes.

4. **Historical Average (HAVG).** Consists of features defined in the baseline plus a feature representing the average number of bicycles at each given time in past days. This feature is only used for long-term predictions.

5. **All Features (ALL).** Consists of the baseline features plus features of the 1st nearest surrounding station, bicycle redistribution activities (last 5 minutes in the case of short-term and mid-term predictions and last 30 minutes in the case of long-term predictions) and historical average (for long-term predictions only).

We name each experiment using the following notation FEATURES-MODEL-DETAILS, e.g. SURR-LR-NEAR10 refers to the experiment of fitting a GAM model with the 10 nearest surrounding stations feature subset.

6.1.1.1 Short-Term Predictions Feature Selection

Figure 6.1 shows the $RMSE_{AVG}$ achieved by the GAM and LR models with the various subsets of features in the short-term prediction scenario. In it we can observe that: most of the subsets produce a similar performance (with the error ranging from 1.026 to 1.076), that the difference between the best performing GAM and the best performing LR model is very small (0.009 bicycles) and that the inclusion of a high number of surrounding stations lowers the GAM's performance, something to what the the LR model seems unaffected.

The summary statistics for the model fitted for station BikePoints_770 (station with median performance) in the ALL-GAM experiment can be observed in listing 2, where the adjusted R-Squared value of 0.992 suggests that the model fits the data well and the p-values table shows that features Near1TMinus2, Near1TMinus3, DistNbBikes, CollNbBikes, Rain and TimeOfDay w/ Holiday have low significance levels (p-values higher than 0.05). The high value for the adjusted R-Squared metric (the amount of variance explained by the model) is misleading if looked on its own, as it must be noted that the BASE-GAM for this same station achieves an equal adjusted R-Squared value using a smaller number of features.

6.1.1.2 Mid-Term Predictions

Figure 6.2 shows the $RMSE_{AVG}$ achieved by our two models with each subset of features in the mid-term prediction scenario. It can be observed that SURR-GAM-

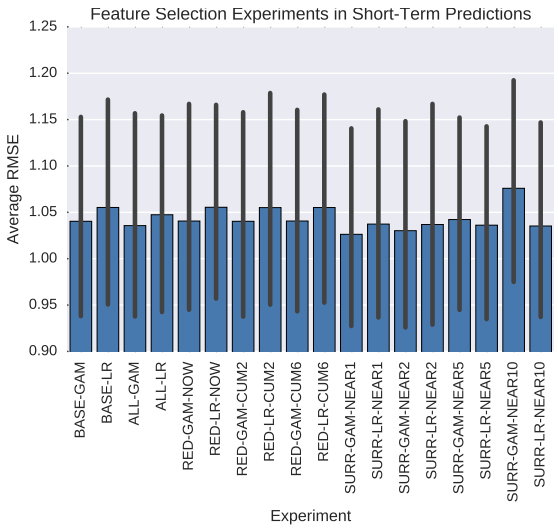

Figure 6.1: Performance of the LR and GAM models with different subsets of features in the short-term prediction scenario. The vertical black lines are error bars set at a confidence interval of 95 %.

NEAR1 achieved the best performance, closely followed by the baseline subset of features. However, the difference between the predictors is minimal, which favors the case of using the simplest subset of features). As in the short-term prediction scenario, a trend involving the number of surrounding stations to include in the GAM can also be observed: the larger the number of surrounding stations, the larger the error. The difference in the performance between the GAM and the LR model (0.096) is still not significant.

As it can be observed in listing 3, which outlines the summary of the fitted model in the ALL-GAM experiment for station with 53th percentile performance (station Id=BikePoints_52), the inclusion of bicycle redistribution features DistNbBikes and CollNbBikes is not helpful to the model as they have low significance levels. The model's explained variance, as measured by the adjusted R-Squared metric), is 83.5 %, however, as in the short-term prediction scenario, a model including only the BASE features explains an almost equal percentage (84.1 %).

6.1.1.3 Long-Term Predictions

Figure 6.3 and listing 4 show the RMSE_{AVG} of each experiment and the summary statistics of the fitted model in the ALL-GAM experiment with 50th percentile performance, respectively. In the former we can observe that LR is the best performing model when used with the ALL set of features (closely followed by HAVG-LR), while

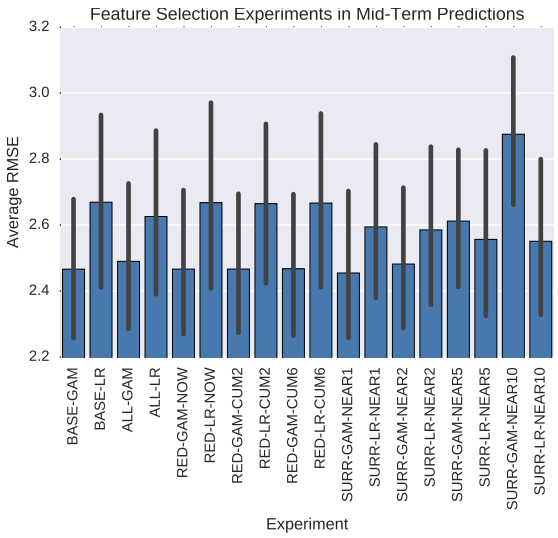

Figure 6.2: Performance of the LR and GAM models using different subsets of features in the mid-term prediction scenario. The vertical black lines are error bars set at a confidence interval of 95 %.

in the latter, that features CollNbBikesCum6, DistNbBikesCum6 and TimeOfDay w/ Holiday have low significance levels. As in the previous prediction scenarios, the explained variance of 12.4 % achieved by this model, can be equaled with a feature set without the redistribution variables. The performance between LR and GAM is significant when using the baseline feature set (1.101 bicycles), however when other features are included, the LR outperforms the GAM, which suggests that the GAM might be overfitting the training data.

6.1.1.4 Selected Feature Sets

Appendix tables B.1, B.2 and B.3 show the exact performance of each feature set in each prediction scenario. While it can be observed that the best performing feature sets were SURR-NEAR1, SURR-NEAR1 and ALL in the short-term, mid-term and long-term prediction scenarios, respectively, we decide to use BASE for the first two scenarios as it achieved almost identical performance as SURR but with a reduced set of features. This same behavior was observed for stations with 1st, 53th and 99th performance percentiles in the three scenarios. For the third scenario we will use ALL as it achieved the best performance.

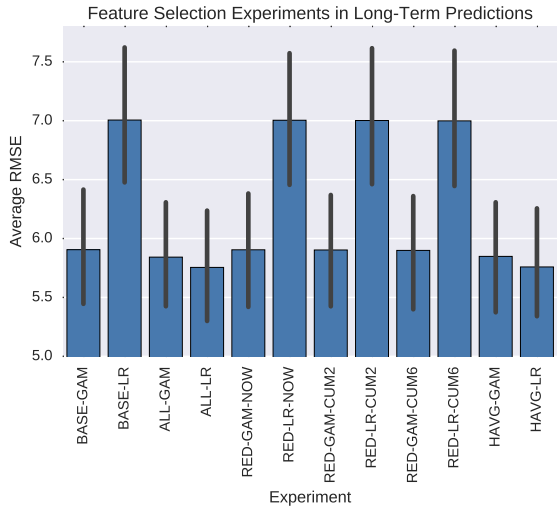

Figure 6.3: Performance of the LR and GAM models using different subsets of features in the long-term prediction scenario. The vertical black lines are error bars set at a confidence interval of 95 %.

	Un-regularized	Regularized	Difference
Short-Term (ALL Feature Set)	0.929181 (0.371)	0.928878 (0.370)	0.000303
Mid-Term (BASE Feature Set)	2.289720 (0.873)	2.318578 (0.893)	-0.028858
Long-Term (HAVG Feature Set)	6.055086 (2.575)	6.049616 (2.579)	0.005470

Table 6.1: Performance of regularized and un-regularized GAM using feature set ALL in each prediction scenario.

6.1.2 Regularization Experiments

Table 6.1 shows the RMSE_{AVG} achieved by regularized and unregularized GAMs for each prediction scenario. In it we can observe that our search to find combinations of smoothing parameters which improved performance significantly failed for each of the prediction scenarios. In the short and long term scenarios, the error decreased by 0.000303 and 0.005470, respectively, while for the mid-term scenario the error actually increased by 0.028858.

6.1.3 Final Predictor Performance

Table 6.2 shows the performance (evaluated using RMSE_{AVG} and $\text{NRMSE}_{\text{AVG}}$) of Hist-AVG, LR and GAM models using the best feature set for each model in each prediction scenario using the best performing feature sets. Surprisingly, the baseline

Metric	Scenario	Hist-AVG	LR	GAM
$RMSE_{AVG}$	Long-Term	**5.400 (1.823)**	5.5190 (1.918)	5.5390 (1.930)
	Mid-Term	5.4020 (1.822)	2.6030 (0.993)	**2.370 (0.858)**
	Short-Term	5.399 (1.8186)	1.040 (0.409)	**1.024 (0.396)**
$NRMSE_{AVG}$	Long-Term	**0.207 (0.037)**	0.211 (0.039)	0.212 (0.039)
	Mid-Term	0.207 (0.0390)	0.104 (0.0399)	**0.095 (0.034)**
	Short-Term	0.207 (0.038)	0.042 (0.018)	**0.041 (0.018)**

Table 6.2: Performance of the models with their best set of features in each prediction scenario.

model, Hist-AVG, performs better than both LR and GAM in the long-term prediction scenario, while in the rest of the scenarios, it is the GAM which achieves performs the best.

6.1.4 Predicted Availability Visualization

We present a geographical information system (GIS) which allows easy visualization of current and predicted bicycle availability throughout the docking stations of Santander Cycles. Our map based solution, shows every docking station as a circle marker, where each marker can take colors ranging from red (station empty) to blue (station full). Further details can be accessed upon selecting a station, which display a pop-up modal with the following information:

1. **Station Name**: The full name of the docking station.
2. **Station Id**: Id of the docking station as identified by TfL's Unified API.
3. **Current Availability**: Current number of available bicycles in the docking station.
4. **Availability 1 Hour Ahead**: Predicted number of bicycles available in the station 1 hour ahead.
5. **Availability 3 Hours Ahead**: Predicted number of bicycles available in the station 3 hours ahead.
6. **Availability Trend**: Plot showing bicycle availability in the station in the past 12 hours.

The map is updated every 5 minutes to reflect the latest availability and predictions. Note that the predictions are done using the best performing model for each prediction scenario as discovered in the previously described experiments.

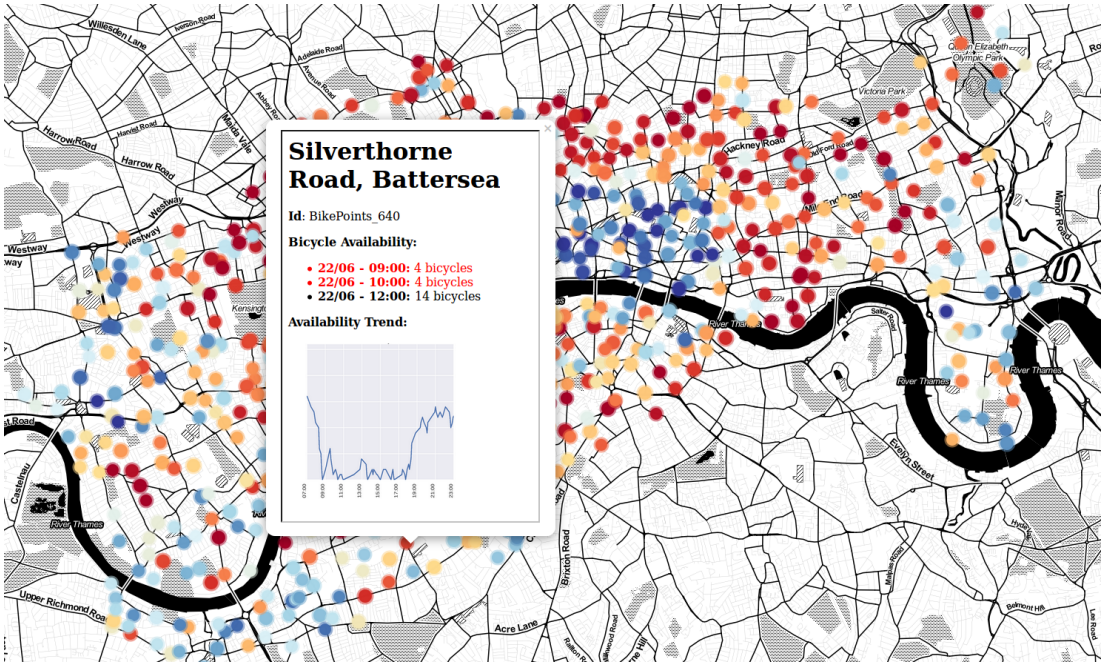

Figure 6.4

6.2 Discussion

The results of the feature selection experiments show that including features that model bicycle availability in up to the 10th nearest surrounding stations has no significant improvement in the performance of the models. A similar result was achieved by Yoon et al [32], who were only able to reduce the error of their ARIMA model slightly (e.g. from a RMSE of 3.50 ± 1.09 to 3.47 ± 1.09) using even more complex neighboring structure methods. These two results contrast with the findings of Yang et al [31], which show a significant performance increase in their Neural Network when using features of the neighboring stations with similar activity patterns. These results lead us to believe first, that more complex models are needed to capture the nonlinear patterns and feature interactions that relate availability in the neighboring stations with the response variable, and second, that the performance increase depends also in the approach used to find the relevant neighboring stations.

Similarly, this same experiments showed that variables that model redistribution

activities in the system did not produce any significant improvements in the predictors' overall performance. However, given that redistribution activities happen rarely (only 0.18 % of the readings have redistribution activity information) and that most redistribution activities are focused on a small number of stations, we believe this could be not because these variables are not significant, but because the evaluation metrics are not precise enough to measure the improvements contributed by these variables. A further experiment using a dataset partitioned in a way such that the proportion of readings with bicycle redistribution activities is larger could be done to assess the significance of these variables.

Our regularization experiments failed to find a combination of smoothing parameters that significantly improves the GAMs' performance. We believe this is because the method used during training to automatically find the smoothing parameters of the splines, Generalized Approximate Cross Validation (GCV), is already achieving good performance due to the large size of our dataset, which allows it to estimate and evaluate the parameters better.

It can also be observed that the GAM and LR models' performance is very similar in the three prediction scenarios, with the smallest difference being 0.016 in the 10 minutes ahead forecast. A closer look at the splines fitted by the GAM for the station with median performance in this scenario (appendix figure B.1) reveals that the splines for features NbBikesTMinus2 and NbBikesTMinus3 show approximately linear patterns and that the spline for the interaction term TimeOfDay/Weekdays has a very large quasi linear section. Given that these three features are among the ones with the highest sGeneralized Cross Validation GCV ignificance (appendix listing 2), it is not surprising that the performance between these models is similar. The splines of the fitted GAMs for the station with median performance in the rest of the prediction scenarios can be observed in appendix figures B.2 and B.3, where the presence of splines with approximately linear behavior can also be frequently observed.

The autocorrelation (ACF) and partial autocorrelation plots (pACF) for the GAM with median performance in each of the three prediction scenarios can be found in figure 6.5. These plots show that there is large residual autocorrelation in the data which the GAMs can not model correctly, the larger the prediction window, the more residual autocorrelation appears. For the short-term scenario, the peak in lag-1 in the ACF plot and the tapering and alternating residuals in the pACF plot are characteristics of a moving average pattern. The pACF plot for the mid-term scenario has spikes at lag-13 and lag-2 5, which given that the readings are spaced evenly every 5 minutes,

correspond to a lag of 1 and 2 hours, respectively. This suggests the presence of an hourly seasonal pattern in the data. In the case of the long-term scenario, the large spike at lag 1 in the pACF means that the lag-1 autocorrelation explains all the higher-order autocorrelations, which could be addressed with a non-seasonal auto-regressive term. While our GAM includes a basic form of auto regressive terms (in the TMinusN terms), these observations suggest that this is not enough to explain the patterns in the data, which favors the case of a more complex model such as a Generalized Additive Mixed Models [7], which is a variant of GAMs that is able to explain both random and fixed effects in the data by supporting correlation structures such as AR and ARMA.

All these observations suggest that a more flexible model that is able to explain complex feature interactions and autocorrelations is needed to achieve better performance when predicting bicycle availability.

While our approach of fitting one model per station has the potential of making the most accurate predictions (as each model is tailored to it's station particular patterns), having a high number of models has several drawbacks. First, training 765 different models is computationally expensive, which made us run the experiments with a sample of 100 stations instead. Second, it is difficult to analyze and diagnose the behavior of such large number of different models. To overcome this issue we look in close detail the models with 1st percentile, median and 99th percentile performance to study the worst, average and best case-scenarios, respectively. And third, storing, loading and using the models proved to be costly too, with our computer equipped with a 6th generation Intel Quad Core i7 Processor, 16 gb of RAM and a Solid State Drive taking almost 10 minutes just to load the file with the 765 models.

6.3 Conclusions

The goal of this project was to use data mining and machine learning to discover valuable insights about the operational and usage patterns of London's bicycle sharing scheme, Santander Cycles, and to predict bicycle availability 10, 60 and 180 minutes ahead in each of it's stations. The increased understanding of the system will enable the implementation of data-driven strategies which make it more efficient, reliable and pleasant to use.

Our work started with a review of previous studies which analyzed and predicted the behavior of bicycle sharing schemes throughout the world. Each of these works identified a set of features that were the most effective when predicting bicycle avail-

ability, which we then combined to establish an unused set of features which, theoretically, when applied to a flexible enough model, would achieve a high prediction accuracy. Such features included data describing time, previous availability readings, weather conditions and surrounding stations availability readings. Furthermore, given that these previous studies suggested that the use of features modeling bicycle redistribution activities could further improve performance but that none of them had actually experimented with them, led us to include them in our experiments. Based on this same literature review, we decided to use a Generalized Additive Model due to it being in the middle ground between an extremely flexible black box model such as a Neural Network and a rigid interpretable model such as Linear Regression.

The data used in our experiments came from various sources such as Transport for London Unified API, Freedom of Information Requests to United Kingdom public authorities and Historical Weather reports. We developed a data ingestion pipeline which ran during 6 weeks in a cloud hosted Linux instance to collect the data for our experiments. The data was then parsed, transformed and loaded into DataFrames for it's storage, manipulation and study. An exploratory data analysis was done for each of the datasets to ensure its' technical correctness and consistency.

Then, to improve our understanding about the usage and operational patterns of the sharing scheme, we first analyzed the station's activity and availability patterns during weekdays and weekends. Where it was discovered that the system displays a typical commute pattern and that it is frequently used as a last mile transportation option. Areas of the city having particularly high docking station activity were also discovered. Furthermore, we studied individual station bicycle availability and bicycle redistribution activities patterns, which helped us gain a deeper understanding of the problems that could arise when modeling the stations' behavior.

Afterwards, the sharing scheme's operational performance was evaluated in our observed 6 week period using the performance indicators regarding bicycle distribution defined in the Service Level Agreement. These performance indicators measure the number of minutes that the docking stations are empty or full and impose a penalization to the service operator when the specified threshold is breached. Here it was discovered that the service operator particularly struggles with preventing stations from becoming full or empty during non-peak hours, disregarding the stations' priority. Also that the failure rates for the performance indicators are much worse during morning peak hours than evening peak hours.

We then focused on predicting bicycle availability in each station of Santander

Cycles using a Historical Average baseline model, a Linear Regression model and a Generalized Additive Model. We defined equations to explain the mathematical relationship between our set of features and the number of bicycles in a station and experiments to select the best performing features using the wrapper method and to regularize our models to prevent them from overfitting. It was discovered that the most significant features are time of day, previous availability readings and weather conditions. Surprisingly, bicycle redistribution and surrounding stations features did not improved the models' performance. We failed to find a combination of smoothing parameters that significantly improves the models' generalization performance, which we believe was caused due the training algorithm ability to automatically find good smoothing parameters when fed with large amounts of data.

Our work concludes by presenting and discussing the results of the carried out experiments, where we present the case that despite the fitted splines modeling the features well, a more complex model is needed as a further inspection of the fitted models revealed unexplained autocorrelations and the need for feature interactions. The results also suggest that a better evaluation strategy is needed to assess the usefulness of the bicycle redistribution features, as we believe that due to the small proportion of records having redistribution data, the improvements are not easily detected when measured along the entire dataset.

6.3.1 Future Work

We believe that future work could be done to assess the impact of bicycle redistribution features with a better evaluation strategy and that further experiments with more complex models and feature interactions could achieve increased performance.

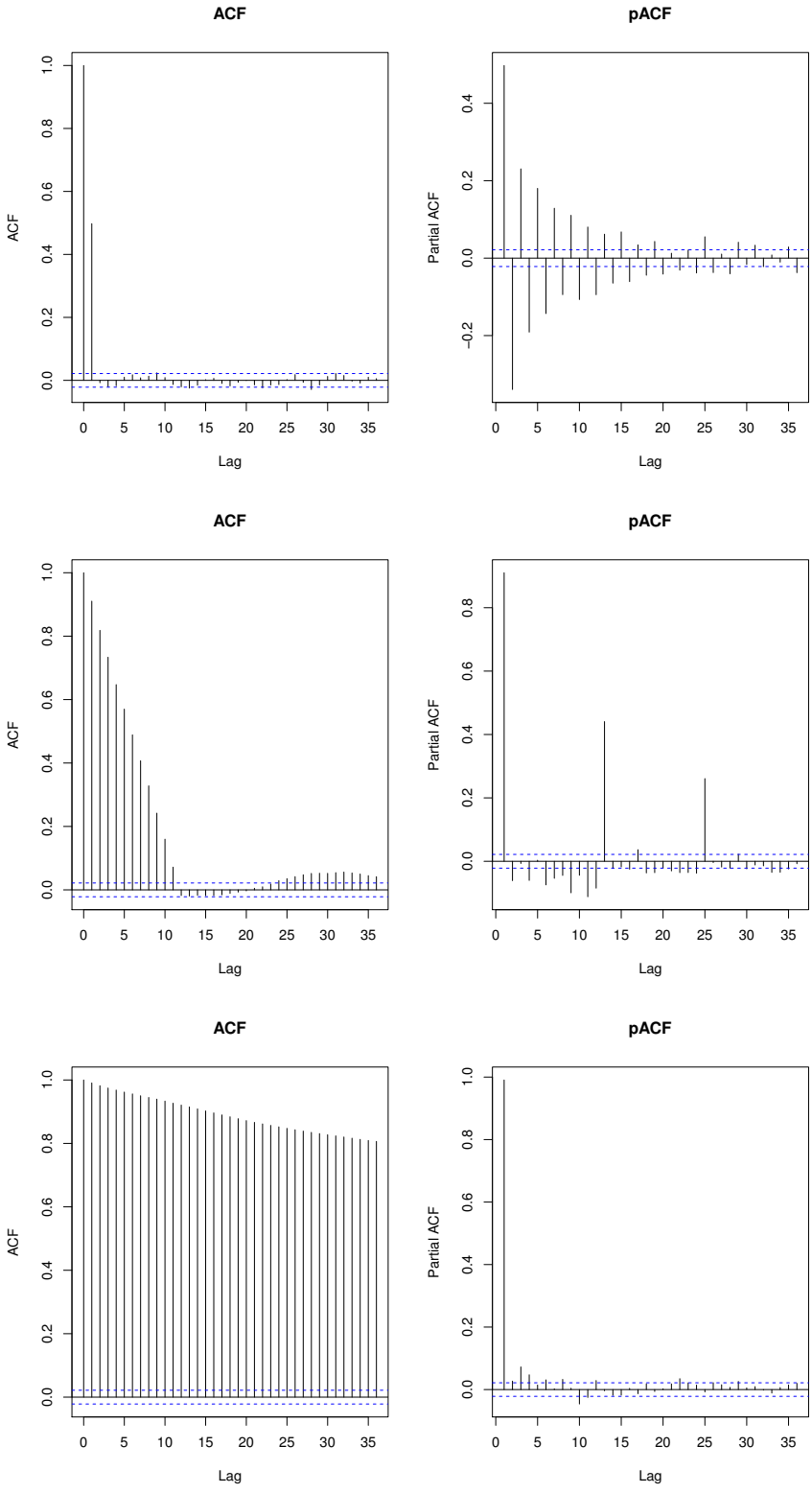

Figure 6.5: Autocorrelation (ACF) and partial autocorrelation plots (pACF) for the short (top), mid (middle) and long (bottom) term GAM models.

Appendix A

Analysis

A.1 Redistribution Activities

Figure A.1: Time of occurrence of bicycle dropped events in the top 12 stations to which most bicycles were distributed. For aesthetics, the used data ranges only from 16 May to 10 June, 2016.

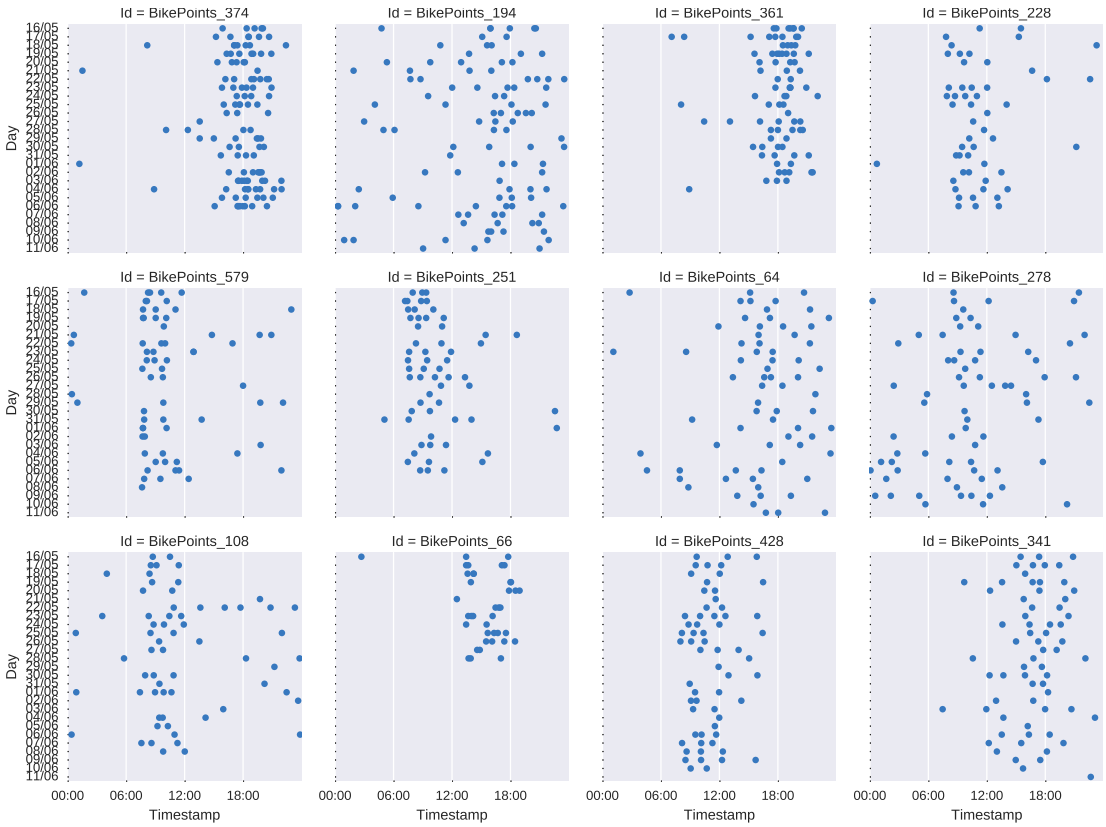

Figure A.2: Time of occurrence of bicycle collection events in the top 12 stations from which most bicycles were collected. For aesthetics, the used data ranges only from 16 May to 6 June, 2016.

A.2 Key Performance Indicators

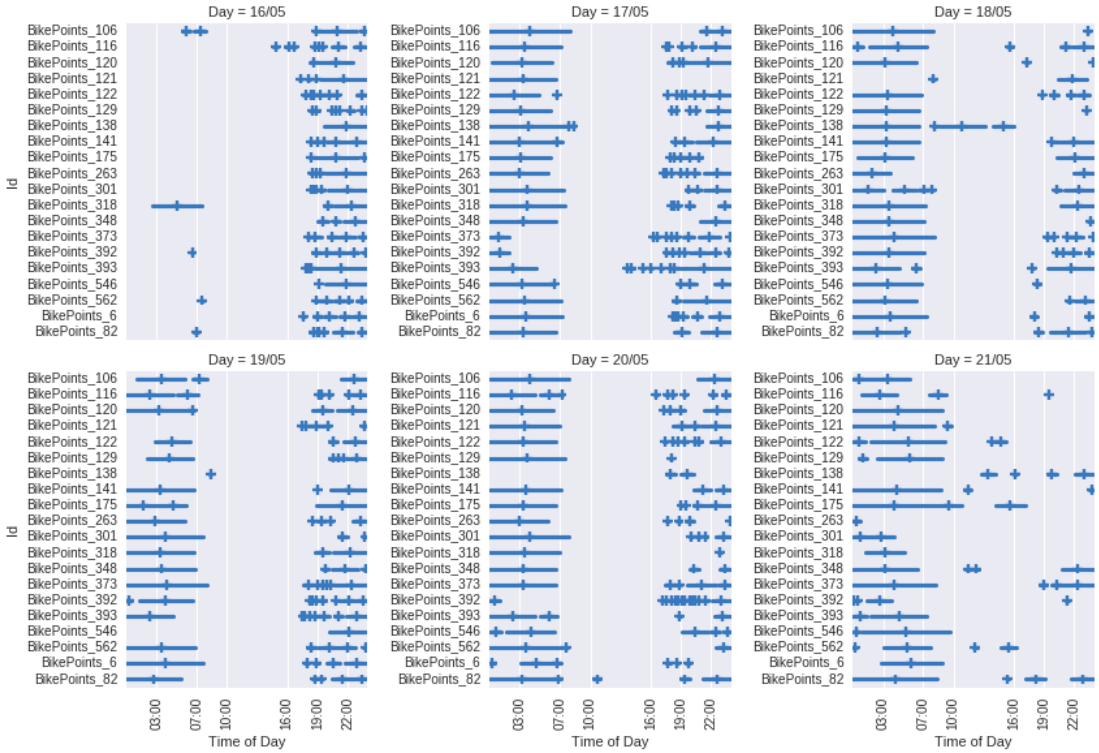

Figure A.3: Empty periods of the 20 stations that accumulated the most empty minutes during the first 6 days of the observational period.

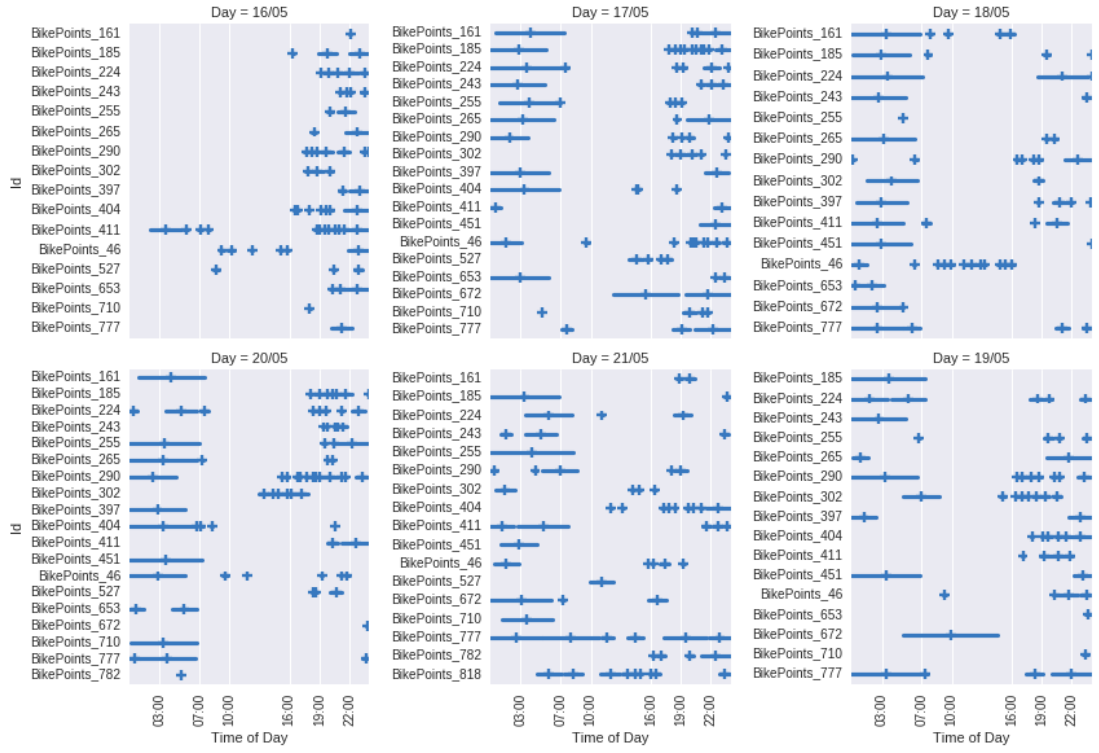

Figure A.4: Full periods of the 20 stations that accumulated the most full minutes during the first 6 days of the observational period.

Figure A.5: Number of accumulated empty and full minutes by P1 and P2 stations during the non-peak hours of each day of the observed 6 week period. The yellow and red lines represent the thresholds of the acceptable number of accumulated minutes for priority 1 and priority 2 stations, respectively.

Figure A.6: Number of accumulated empty and full minutes by priority 1 stations during morning and evening peak time hours. The red line represents the acceptable service level threshold.

Figure A.7: Number of accumulated empty and full minutes by priority 2 stations during morning and evening peak time hours. The red line represents the acceptable service level threshold.

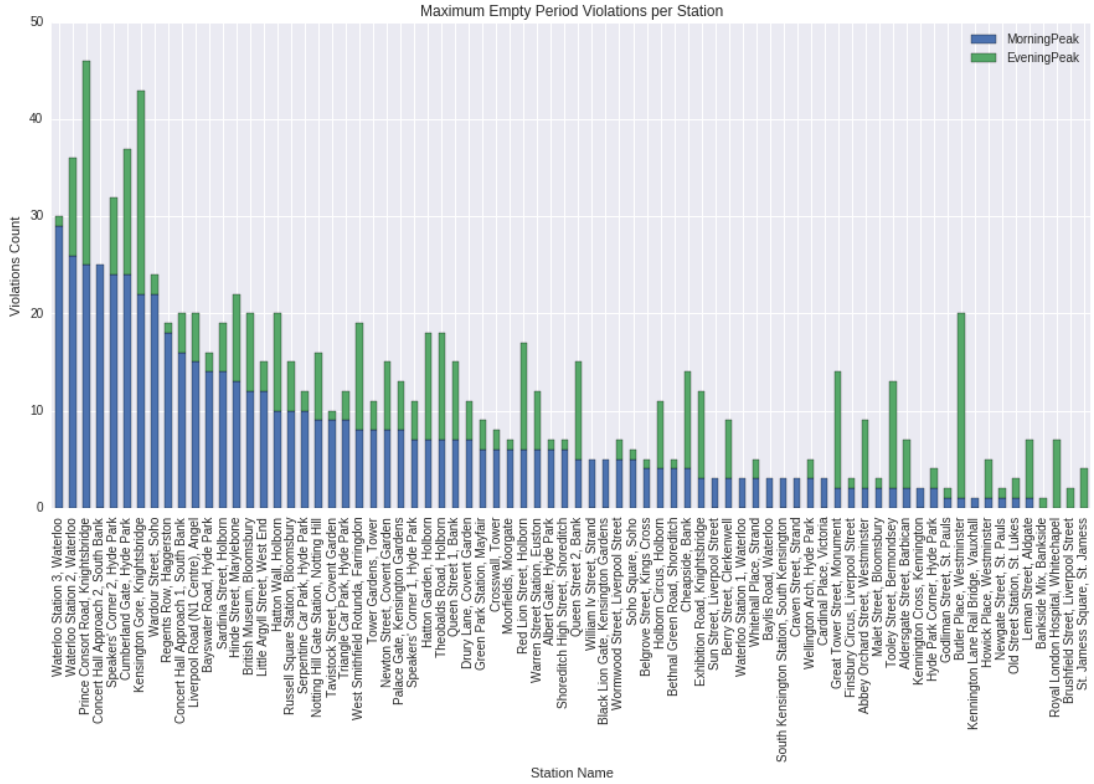

Figure A.8: Number of violations of PI 26 per station.

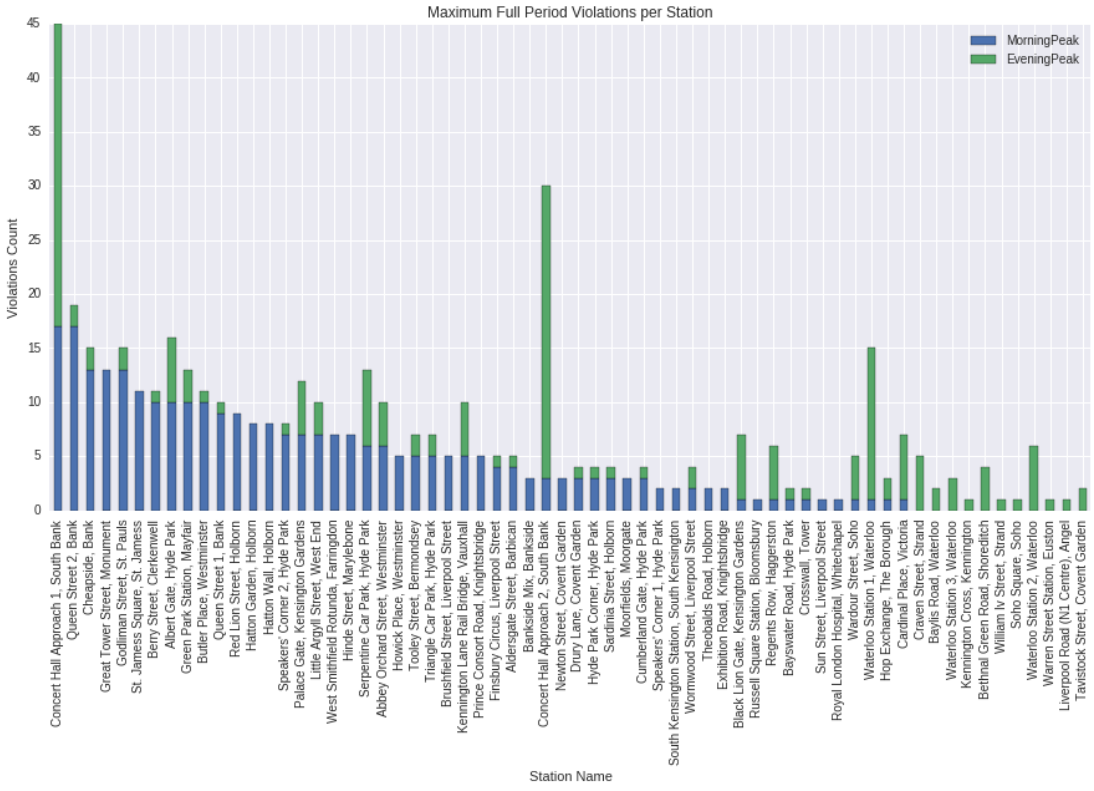

Figure A.9: Number of violations of PI 27 per station.

Appendix B

Results and Conclusions

Feature Set	$RMSE_{AVG}$
SURR-GAM-NEAR1	1.026 (0.560)
SURR-GAM-NEAR2	1.030 (0.560)
SURR-LR-NEAR10	1.035 (0.572)
ALL-GAM	1.036 (0.559)
SURR-LR-NEAR5	1.036 (0.571)
SURR-LR-NEAR2	1.037 (0.570)
SURR-LR-NEAR1	1.037 (0.572)
RED-GAM-CUM2	1.040 (0.566)
BASE-GAM	1.040 (0.567)
RED-GAM-CUM6	1.041 (0.566)
RED-GAM-NOW	1.041 (0.567)
SURR-GAM-NEAR5	1.042 (0.563)
ALL-LR	1.047 (0.571)
RED-LR-CUM6	1.055 (0.579)
RED-LR-CUM2	1.055 (0.579)
RED-LR-NOW	1.055 (0.580)
BASE-LR	1.055 (0.580)
SURR-GAM-NEAR10	1.076 (0.564)

Table B.1: Performance of each feature set for the short-term prediction scenario.

Feature Set	$RMSE_{AVG}$
BASE-GAM	2.466 (1.122)
BASE-LR	2.669 (1.318)
ALL-GAM	2.490 (1.106)
ALL-LR	2.626 (1.269)
RED-GAM-NOW	2.466 (1.122)
RED-LR-NOW	2.668 (1.316)
RED-GAM-CUM2	2.467 (1.122)
RED-LR-CUM2	2.665 (1.313)
RED-GAM-CUM6	2.467 (1.122)
RED-LR-CUM6	2.667 (1.317)
SURR-GAM-NEAR1	2.455 (1.105)
SURR-LR-NEAR1	2.594 (1.273)
SURR-GAM-NEAR2	2.482 (1.102)
SURR-LR-NEAR2	2.585 (1.257)
SURR-GAM-NEAR5	2.612 (1.125)
SURR-LR-NEAR5	2.557 (1.235)
SURR-GAM-NEAR10	2.875 (1.187)
SURR-LR-NEAR10	2.551 (1.214)

Table B.2: Performance of each feature set for the mid-term prediction scenario.

Feature Set	$RMSE_{AVG}$
BASE-GAM	5.905 (2.422)
BASE-LR	7.006 (2.905)
ALL-GAM	5.842 (2.425)
ALL-LR	5.755 (2.429)
RED-GAM-NOW	5.904 (2.423)
RED-LR-NOW	7.004 (2.905)
RED-GAM-CUM2	5.903 (2.424)
RED-LR-CUM2	7.002 (2.905)
RED-GAM-CUM6	5.899 (2.426)
RED-LR-CUM6	6.999 (2.908)
HAVG-GAM	5.848 (2.420)
HAVG-LR	5.759 (2.425)

Table B.3: Performance of each feature set for the long-term prediction scenario.

```
Family: gaussian
Link function: identity

Formula:
NbBikes ~ s(TempTMinus2, HumidityTMinus2, bs = "tp") + s(TimeOfDay,
    by = Weekday, bs = "tp") + s(TimeOfDay, by = Weekend, bs = "tp") +
    s(TimeOfDay, by = Holiday, bs = "tp") + s(NbBikesTMinus2,
    bs = "tp") + s(NbBikesTMinus3, bs = "tp") + RainTMinus2 +
    FogTMinus2 + s(Near1TMinus2, bs = "tp") + s(Near1TMinus3,
    bs = "tp") + CollNbBikes + DistNbBikes

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.768022   0.153816  70.006   <2e-16 ***
RainTMinus2 -0.033561   0.029458  -1.139    0.255
FogTMinus2   0.000000   0.000000      NA       NA
CollNbBikes  0.003795   0.046846   0.081    0.935
DistNbBikes  0.037832   0.033490   1.130    0.259
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                                  edf Ref.df       F  p-value
s(TempTMinus2,HumidityTMinus2) 20.107 24.683   1.777   0.0103 *
s(TimeOfDay):Weekday            9.516  9.659 195.958  < 2e-16 ***
s(TimeOfDay):Weekend            8.532  9.331 185.657  < 2e-16 ***
s(TimeOfDay):Holiday            2.000  2.000   3.600   0.0274 *
s(NbBikesTMinus2)               5.763  6.744 651.132  < 2e-16 ***
s(NbBikesTMinus3)               6.622  7.460   7.093 1.27e-08 ***
s(Near1TMinus2)                 7.631  8.305   2.243   0.0224 *
s(Near1TMinus3)                 6.239  7.349   1.352   0.2378
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Rank: 98/100
R-sq.(adj) =  0.992   Deviance explained = 99.2%
GCV = 0.4956  Scale est. = 0.49136   n = 8183
```

Listing 2: Summary statistics for the GAM fitted for station BikePoints_770 for the short-term prediction scenario.

```
Family: gaussian
Link function: identity

Formula:
NbBikes ~ s(TempTMinus12, HumidityTMinus12, bs = "tp") + s(TimeOfDay,
    by = Weekday, bs = "tp") + s(TimeOfDay, by = Weekend, bs = "tp") +
    s(TimeOfDay, by = Holiday, bs = "tp") + s(NbBikesTMinus12,
    bs = "tp") + s(NbBikesTMinus18, bs = "tp") + s(Near1TMinus12,
    bs = "tp") + s(Near1TMinus18, bs = "tp") + CollNbBikes +
    DistNbBikes

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.10854    0.47890  19.020   <2e-16 ***
CollNbBikes -0.02302    0.08586  -0.268    0.789
DistNbBikes  0.12944    0.14817   0.874    0.382
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                                    edf Ref.df       F  p-value
s(TempTMinus12,HumidityTMinus12) 25.951 28.383   7.516  < 2e-16 ***
s(TimeOfDay):Weekday              9.649  9.666 245.406  < 2e-16 ***
s(TimeOfDay):Weekend              9.194  9.602  19.528  < 2e-16 ***
s(TimeOfDay):Holiday              9.903  9.997  63.861  < 2e-16 ***
s(NbBikesTMinus12)                8.379  8.881 295.568  < 2e-16 ***
s(NbBikesTMinus18)                7.911  8.680   5.223 9.72e-07 ***
s(Near1TMinus12)                  8.806  8.982  18.077  < 2e-16 ***
s(Near1TMinus18)                  7.025  8.105   8.256 2.89e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Rank: 97/98
R-sq.(adj) =   0.84   Deviance explained = 84.1%
GCV = 4.1386  Scale est. = 4.0933    n = 8168
```

Listing 3: Summary statistics for the GAM fitted for station BikePoints_52 for the mid-term prediction scenario.

```
Family: gaussian
Link function: identity

Formula:
NbBikes ~ s(TimeOfDay, by = Weekday, bs = "tp") + s(TimeOfDay,
    by = Weekend, bs = "tp") + s(TimeOfDay, by = Holiday, bs = "tp") +
    s(HistAvg, bs = "tp", k = 5) + CollNbBikesCum6 + DistNbBikesCum6

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.14084    0.53005   9.699   <2e-16 ***
CollNbBikesCum6 -0.09851    0.06566  -1.500    0.134
DistNbBikesCum6  0.13594    0.07686   1.769    0.077 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                       edf Ref.df      F p-value
s(TimeOfDay):Weekday 8.422  9.264  6.100 5.7e-09 ***
s(TimeOfDay):Weekend 5.681  6.780  3.283 0.00208 **
s(TimeOfDay):Holiday 2.000  2.000  3.647 0.02612 *
s(HistAvg)           1.000  1.000 83.421 < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Rank: 36/37
R-sq.(adj) =  0.122   Deviance explained = 12.4%
GCV = 12.231  Scale est. = 12.201    n = 8186
```

Listing 4: Summary statistics for the GAM fitted for station BikePoints_240 for the long-term prediction scenario.

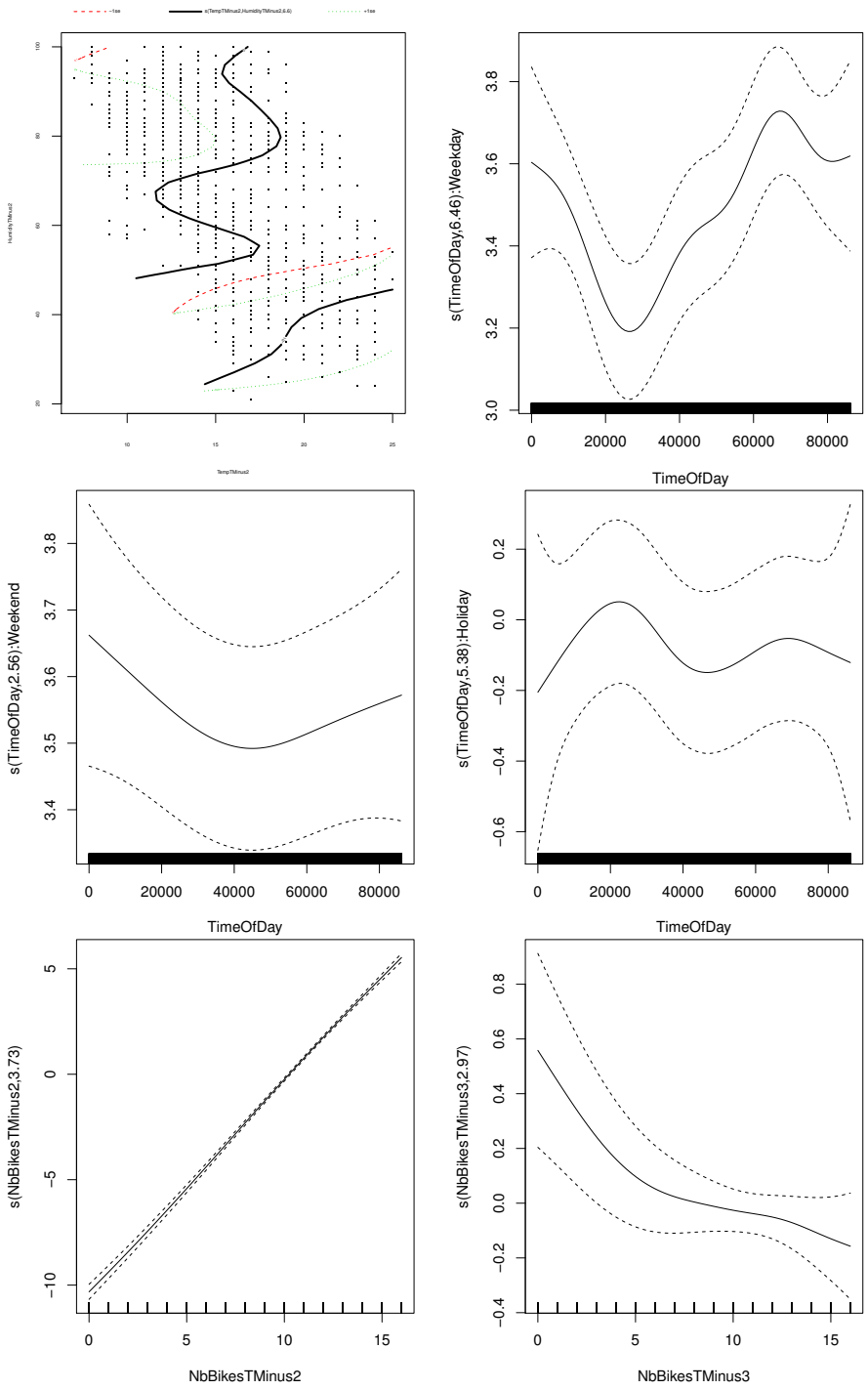

Figure B.1: Splines fitted by the median performing gam for the short-term prediction scenario. They correspond to the smooth functions of the feature interaction TempTMinus2 and HumidityTMinus2 and features TimeOfDay/Weekday, TimeofDay/Weekend, TimeOfDay/Holiday, NbBikesTMinus2 and NbBikesTMinus3.

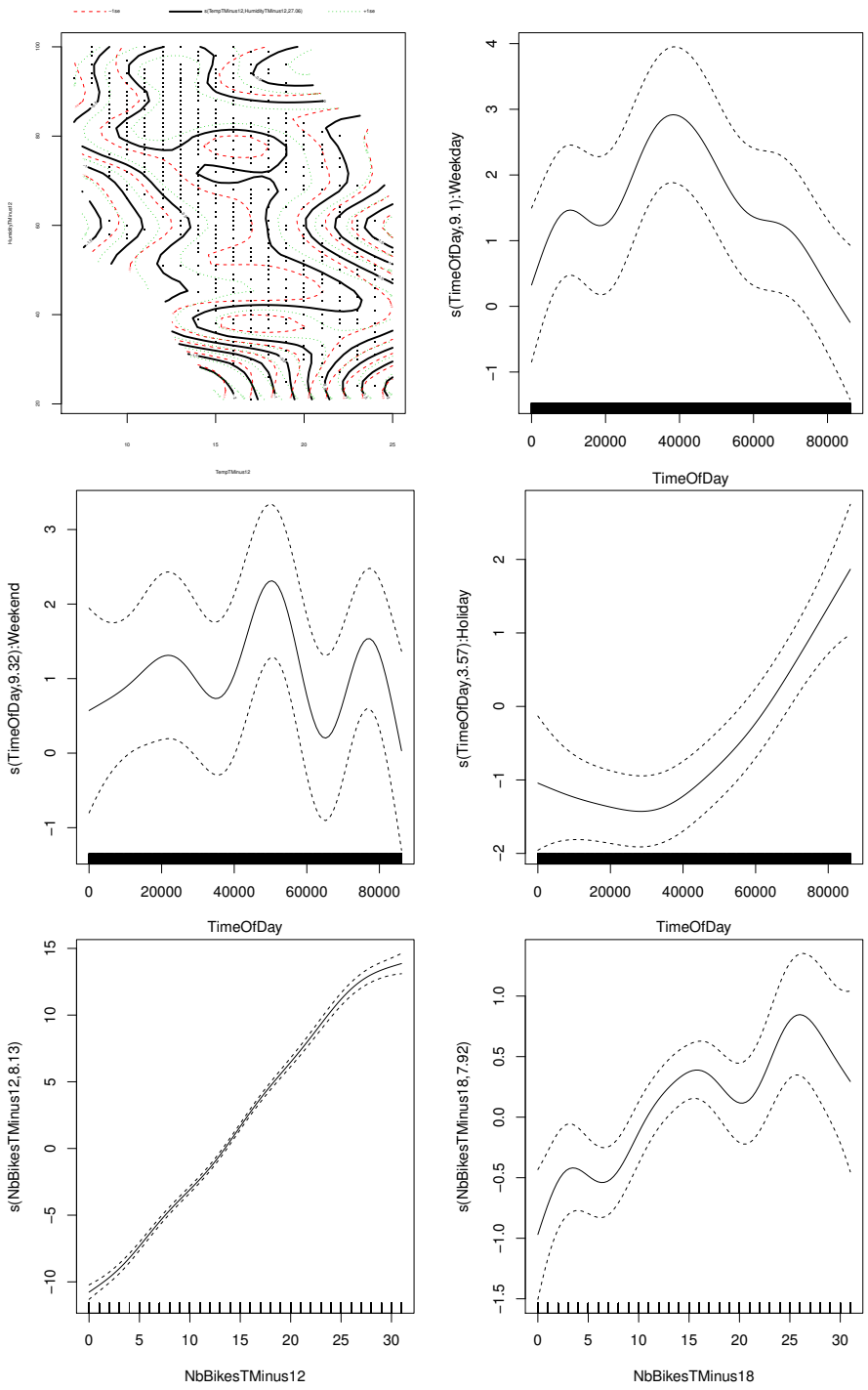

Figure B.2: Splines fitted by the median performing gam for the mid-term prediction scenario. They correspond to the smooth functions of the feature interaction TempTMinus12 and HumidityTMinus12 and features TimeOfDay/Weekday, TimeofDay/Weekend, TimeOfDay/Holiday, NbBikesTMinus12 and NbBikesTMinus18.

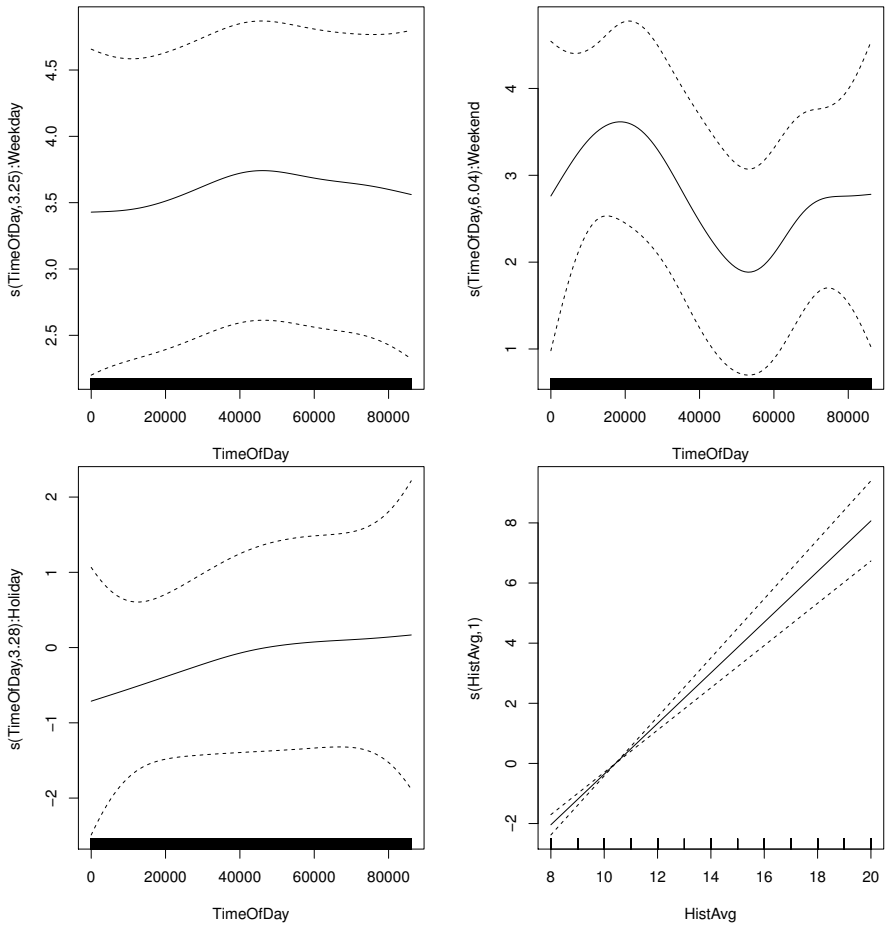

Figure B.3: Splines fitted by the median performing gam for the mid-term prediction scenario. They correspond to the smooth functions of features TimeOfDay/Weekday, TimeofDay/Weekend, TimeOfDay/Holiday and HistAvg.

Bibliography

[1] Walk time calculator. *Walk Time Calculator*, 2016. http://www.walkingenglishman.com/walktime.aspx.

[2] Greater London Authority. Gla 2014 round of trend-based population projections - results. 2015. https://files.datapress.com/london/dataset/2014-round-population-projections/update-03-2015-2014rnd-trend-proj-results.pdf.

[3] 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.

[4] RF CASEY, LN Labell, L Moniz, JW Royal, M Sheehan, T Sheehan, A Brown, M Foy, M Zirker, Carol L Schweiger, et al. Advanced public transportation systems: The state of the art update 2000. Technical report, 2000.

[5] Bei Chen, Fabio Pinelli, Mathieu Sinn, Adi Botea, and Francesco Calabrese. Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots. In *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*, pages 53–58. IEEE, 2013.

[6] Sabeur Elkosantini and Saber Darmoul. Intelligent public transportation systems: A review of architectures and enabling technologies. In *Advanced Logistics and Transport (ICALT), 2013 International Conference on*, pages 233–238. IEEE, 2013.

[7] Ludwig Fahrmeir and Stefan Lang. Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220, 2001.

[8] Lino Figueiredo, Isabel Jesus, JA Tenreiro Machado, J Ferreira, and JL Martins De Carvalho. Towards the development of intelligent transportation systems. In *Intelligent Transportation Systems*, volume 88, pages 1206–1211, 2001.

[9] Transport for London. London cycle hire service agreement schedule 5 service level agreement. 2009. `http://content.tfl.gov.uk/lchs-schedule05-service-level-agreement-redacted.pdf`.

[10] Transport for London. Santander cycles customer satisfaction and usage survey: Members only: Wave 11 (quarter 3 2015/16). 2015. `http://content.tfl.gov.uk/bch-members-q3-2014-15.pdf`.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[12] Jon Froehlich, Joachim Neumann, and Nuria Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI*, volume 9, pages 1420–1426, 2009.

[13] Romain Giot and Raphaël Cherrier. Predicting bikeshare system usage up to one day ahead. In *Computational intelligence in vehicles and transportation systems (CIVTS), 2014 IEEE symposium on*, pages 22–29. IEEE, 2014.

[14] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical science*, pages 297–310, 1986.

[15] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

[16] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010.

[17] Russell Meddin and Paul DeMaio. The bike-sharing world map, 2009.

[18] Peter Midgley. Bicycle-sharing schemes: enhancing sustainable mobility in urban areas. *United Nations, Department of Economic and Social Affairs*, pages 1–12, 2011.

[19] Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[20] United Nations. Department of Economic and Social Affairs. Population Division. *World Urbanization Prospects: The 2014 Revision*. UN, 2014.

[21] Mayor of London. The mayor's vision for cycling in london. March 2013. `http://content.tfl.gov.uk/gla-mayors-cycle-vision-2013.pdf`.

[22] Mayor of London. Travel in london report 8. 2015. `http://content.tfl.gov.uk/travel-in-london-report-8.pdf`.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] Parliament Hill Research. 14 cost eective actions to cut central london air pollution. June 2012. `https://www.rbkc.gov.uk/pdf/air_quality_cost_effective_actions_full_report.pdf`.

[25] Ford Rylander, Bo Peng, and Jeff Wheeler. Bike share usage prediction in london. Unpublished work, 2014.

[26] Tom Thomas, Rinus Jaarsma, and Bas Tutert. Exploring temporal fluctuations of daily cycling demand on dutch cycle paths: the influence of weather on cycling. *Transportation*, 40(1):1–22, 2013.

[27] Department for Business Innovation & Skills; Uk. Smart Cities: Background paper. (October):47, 2013.

[28] Hadley Wickham et al. Tidy data. *Under review*, 2014.

[29] Simon Wood. *Generalized additive models: an introduction with R*. CRC press, 2006.

[30] Min Yang, Yingnan Guang, and Xuedan Zhang. Public bicycle prediction based on generalized regression neural network. In *Internet of Vehicles-Safe and Intelligent Mobility*, pages 363–373. Springer, 2015.

[31] Min Yang and Xuedan Zhang. A novel travel adviser based on improved back-propagation neural network. In *2016 7th International Conference on Intelligent Systems, Modelling and Simulation*. 2016.

[32] Ji Won Yoon, Fabio Pinelli, and Francesco Calabrese. Cityride: a predictive bike sharing journey advisor. In *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, pages 306–311. IEEE, 2012.