©Copyright 2016
Weiran Zhao

Modeling Seasonal and Weather Impacts on Cycling Count

Weiran Zhao

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Urban Planning

University of Washington

2016

Reading Committee:

C.-H Christine Bae, Chair

Don MacKenzie

Program Authorized to Offer Degree:
Department of Urban Design and Planning

University of Washington

Abstract

Modeling Seasonal and Weather Impacts on
Cycling Count

Weiran Zhao

Chair of the Supervisory Committee:
C.-H Christine Bae
Department of Urban Design and Planning

Cycling has been proven to contribute to not only cyclists' heath but also a sustainable transportation system in urban cities. Policy makers and urban planners all over the world have been promoting bicycling. City of Seattle has been implementing new policies and programs to create a bike friendly environment and aims to quadruple ridership by 2030. Therefore, empirically confirming such growth in Seattle will help to justify current and future investment in bicycle infrastructure and programs in Seattle. This study uses Seattle cyclist count data to quantify cycling trend and examine their relationship between seasonal and weather factors. First, a systematic approach is taken to identify the explanatory variables and their appropriate forms of transformation. Then different models are investigated and compared to best capture the relationship between bike counts and factors. Specifically, non-linearity, discontinuity and interaction items are taken into account. Results are interpreted with intuitive visualization using counterfactual simulations. Furthermore, a predictive model is proposed to estimate daily count in the future. Autoregressive Integrated Moving Average (ARIMA) model is used to account for autocorrelation. Its predictive performance is evaluated using cross validation. Finally, proposed methodology is applied to multiple locations in Seattle and identify their unique bicycle travel patterns. This research will help policy makers and transportation planners to better understand the factors that could drive the bike demand and influence bike travel behavior.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables v

Chapter 1: Introduction 1
1.1 Background 1
1.2 Objectives and Contributions of this Thesis 4

Chapter 2: Literature Review 7
2.1 Cycling Ridership Study 7
2.2 Weather and Seasonal Impacts on Bike Count 8
2.3 Modeling Methods and Prediction 10
2.4 Limitations of Current Literature 12

Chapter 3: Methodology 14
3.1 Data Collection, Processing and Description 14
3.2 Descriptive Analysis and Variable Selection 17
3.3 Modeling Approach 26

Chapter 4: Weather and Seasonal Impacts 36
4.1 Generalized Additive Mixture (GAM) Results 36
4.2 Weather Impact 38
4.3 Seasonal Impact 42
4.4 Interacting Relationship 47
4.5 Trend Analysis 49
4.6 Master Model 52

Chapter 5: Predictive Model 54
5.1 Regression-based Model 54
5.2 Time Series Analysis with ARIMA 55

Chapter 6: Multi-Location Analysis 61
6.1 Locations Overview 61
6.2 Model Results 62

Chapter 7: Conclusion 71

Bibliography 74

LIST OF FIGURES

Figure Number Page

3.1 Location of Fremont Bridge 15
3.2 EcoCounter on the Fremont Bridge 16
3.3 Descriptive visualization of bicycle counts dataset: (a) Frequency of observed counts; (b) Timeseries of counts. 18
3.4 Scatter plot Matrix of covariates in the Group of Temperature. 21
3.5 Scatterplot Matrix of covariates in the Group of Precipitation. 22
3.6 Count box plots of `dow`, `Weather`, `UW`, `Weekend`, `Holiday`. 24
3.7 Residual plots for different types of models 30
3.8 Box-Cox test 32

4.1 Smooth functions of `TempMax` and `PrecipProb` 38
4.2 Effect of `TempMax` on bicycle counts, all other factors held constant, with shaded 95% confidence interval. 40
4.3 Effect of `PrecipProb` on bicycle counts, all other factors held constant, with shaded 95% confidence interval. 41
4.4 Effect of `Daylight` and `UW` in-session status on bicycle counts, all other factors held constant, with shaded 95% confidence interval. 44
4.5 Effect of `weekend` on bicycle counts, all other factors held constant, with 95% confidence interval. 45
4.6 Effect of `Season` on bicycle counts, all other factors held constant, with 95% confidence interval. 46
4.7 Effect of `holiday` on bicycle counts, all other factors held constant, with 95% confidence interval. 46
4.8 Effect of `dow` on bicycle counts, all other factors held constant, with 95% confidence interval. 49
4.9 General trend in bicycling counts, all other factors held constant, with shaded 95% confidence interval. 51

5.1 Actual vs. predicted daily bike volume for each of the four models in Table 5.1. The red line shows actual count is equal to predicted. This plot shows the majority variation is predicted, with some points under-predicted. 56
5.2 From top to bottom: the residual plots, ACF and PACF of the square root transformation model with ARIMA(1,0,1) error terms (right) and without ARIMA error terms (left). 58
5.3 Plot of actual and predicted counts for 2015 test data with 95% predcition limits. . 60

6.1 Five bike counter facilities on map 63
6.2 Bike counts for five locations 64
6.3 Distribution of bike counts at 5 locations 65
6.5 Effect of precipitation (a) and temperature (b) on counts for five locations, with shaded 95% confidence interval. 68
6.6 Effect of weekend on counts for five locations, with 95% confidence interval. . . . 68
6.7 Effect of holiday on counts for five locations, with 95% confidence interval. 69
6.8 The impact of `UW` in session on counts for 5 locations, with 95% confidence interval. 69
6.9 The Impact of daylight hour on counts for 5 locations, with shaded 95% confidence interval. 70
6.10 The Impact of calendar-based seasons on counts for 5 locations. 70

LIST OF TABLES

Table Number Page

3.1 Goodness-of-fit results of models with different forms of temperature and precipitation variables, all other variables controlled. 20
3.2 Candidate explanatory variables 25
3.3 Four types of models 28
3.4 Goodness-of-fit results of models 31

4.1 Model specifications 37
4.2 Results: Weather variables 43
4.3 Results: Seasonal factors 48
4.4 Results: Interaction and trend 50
4.5 Results: Master model 53

5.1 Predication performance on test data set in terms of RMSE, using the first two years' data as training data set 55
5.2 Statistical stationarity test with its $p-$value and null hypothesis 55
5.3 Square root transformation model with vs. without ARIMA error terms 59

6.1 Five Locations with bike counters 62
6.2 Location Comparison 67

Chapter 1

INTRODUCTION

1.1 Background

Cycling is a fun and healthy choice of transportation in urban city environments which can be used for all ages and abilities. It is becoming an increasingly important mode in a comprehensive sustainable transportation system. A country-wide increase in bicycling in urban cities has been seen recently. Policy makers and urban planners all over the world have been promoting bicycling in order to create bikeable cities. Thus, understanding and quantifying bike ridership is critical to planners in the further.

A bikeable environment has been proven to have the following benefits. First, it benefits the health of those choosing to bicycle regularly. Active transportation including cycling has been promoted by public health professions as a means to improve health conditions like cardiovascular disease, diabetes, other chronic diseases, and especially to improve children obesity epidemic. Most cities have Safe Routes to Schools Programs within their transportation departments to encourage kids walking and biking to school [36, 12]. Second, riding bike instead of driving cars could offer substantial economic benefits in urban areas by reducing traffic congestion and energy consumption [57]. More importantly, it helps build low-carbon and green-growth urban environments by reducing gas emissions, thus mitigating climate change and global warming issues. Furthermore, improved safety of all roadway users has been shown in studies. For example, risk of injury or death in a collision with motor vehicles declines as more people walk or bicycle [62, 66]. Bike friendly road design also helps drivers to drive with more caution. Additionally, biking could promote transport equity, enhance social cohesion and community livability [38] as it represents a more affordable and accessible alternative to automobiles.

As such, City of Seattle has been making increased efforts to promote bicyling in the past

5 years. The Seattle Bicycle Master Plan (BMP) was passed in 2014 to set a detailed blueprint to design and implement bicycle facilities that are safe and appropriate for riders of all ages and abilities. It expanded bicycle infrastructure (e.g., cycle tracks, bicycle lanes, bicycle parking) as well as implementing new policies and programs (e.g. bike sharing programs, traffic calming, bicycle integration with transit). The updated BMP of 2015 aims to a quadruple ridership with a full bike facilities coverage in the City by 2030. The plan recommends that 474 new or upgraded facilities are to be added to the current existing 135 networks. The current cost estimate of full build-out of the bicycle facility projects in Seattle ranges from $390-$524 million [3].

With initiatives of this size, an increase in cycle ridership has been observed in the last two years and will be continuing in the future. Therefore, empirically confirming such growth in Seattle will help justify current and future investment in bicycle infrastructure and programs in Seattle. Reliable bike count estimations are essential to determine whether current corridor designs are working well and for future design of appropriate infrastructure that can accommodate bicycles, pedestrians and vehicles safely. Moreover, in order to develop policy and improve relevant infrastructure planning to induce more bicycling, it is becoming more and more critical to understand the determinants of bicycle ridership and their dynamic relationships with temperal factors.

Some of the identified determinants include physical, demographic, social-economical and cultural factors [67, 49, 33], accessibility to bicycle infrastructure [14], land use pattern and infrastructure [15]. Although there have been quite rich literature on the bicycle usage, the vast majority of them are focused on topics such as road safety analysis [32], impact of built environment [49], socio-demographic factors and policy [19, 67]. There is relatively little research that investigates the impacts of weather variables and seasonal factors on bicycle ridership. Most research that examined weather factors is limited to automobiles studies [41, 53, 45, 47]. Of the limited studies that do focus on the relationship between bike ridership and weather/seasonal variations, many used manually collected bicycle counts or self-reported survey data based on cyclists' perceptions [42, 65, 51, 52] to determine how weather affects people's decision to cycle. Such approach naturally introduces inaccuracy to the result and fails to quantitatively determine how actual bike flows respond to weather conditions.

One of the major reasons for the limited number of work on bike ridership modeling may be due to the lack of data. Such lack of data is two-fold: the lack of bicycle count in general and the lack of continuously collected count data. Transportation planning in the US has traditionally focused on automotive traffic but is increasingly turning towards a multi-modal approach in order to accommodate all users. States and municipalities are tasked with annually counting the number of motor vehicles traveling their roads through the federally mandated Highway Performance Monitoring System. Unfortunately, few states and municipalities have formal procedures for counting bicyclists. Without a federal mandate, most agencies forego tracking non-motorized forms of travel, other than perhaps conducting a limited assortment of spot-counts at intersections or trail segments [16]. This approach can result in substantial unmet needs and inappropriate investments and policies, as decision-makers rely on minimal demand data for bike routes. On the other hand, the traditional bike count data are usually collected manually over a short-time period because of labor constraints. However, to adequately describe the effect of weather conditions on bike volumes, continuous automated collection of data over a long span of time is necessary [41]. In 2011 SDOT began a new systematic bicycle counts program, with the advent of automated bike counters that are now installed in multiple sites across the City of Seattle, now there is sufficient data to investigate the relationship between bike ridership and weather conditions and other seasonal factors.

A robust understanding of the relationship between bicycle volume and various factors can be helpful to policy makers, urban planners, and researchers in many aspects [2], including: (a) determine existing travel patterns and demand; (b) identify corridors where current use and potential for increased use is high; (c) track trends over time; (d) evaluate the effectiveness of programs and/or facilities to promote biking; (e) identify locations for bicycle facility improvements and design appropriate treatments; (f) measure demographic changes as facilities that increase user comfort and attract a wider range of pedestrians and bicyclists are developed; (g) assess future bicycle travel demand; (h) make informed transportation decisions to prioritize bicycle improvement projects; and (i) appropriately allocate resources for active transportation.

1.2 Objectives and Contributions of this Thesis

To contribute to the small body of research literature investigating the impact of weather conditions and seasonal factors on cycle ridership, there are three major objectives in this thesis:

1. **Provide a thorough study of the impact of different weather variables and seasonal factors on daily bike volume.** A systematic approach is adopted to select variables that have most significant impact on bike ridership. Regression model is used to investigate many different representations and measurements of weather and seasonal factors. Non-linearity, discontinuity as well as interaction relationship are examined and interpreted as well.

2. **Develop a predictive model to estimate daily bike volume.** A good prediction of bike ridership helps transportation administrator to make better-informed preparations/decisions in case of inclement weather or special holidays that would result in significant change in bike counts. The predictive model is proposed by using time series model to accommodate autocorrelation in observations.

3. **Quantify the real bicycle trip trend after excluding the effect of weather and other temporal/seasonal variations.** This study is useful to understand ridership at the aggregated level and help determine the current travel patterns/demand. It will also help to reveal multiple aspects of planning implications such as justify the effectiveness of current and future investments in facilities to promote biking.

The main contributions of this thesis are as follows:

1. **Identify key explanatory variables for the daily bike counts**. A systematic approach is used to select key weather and seasonal factors. Rationale of selection is based on examining studies in previous research and availability of data. Model performance using different forms of variables is considered in order to select proper explanatory variables.

2. **Propose a regression model that can adequately capture the relationship between various factors and bike counts**.

(1) Propose the proper transformation of dependent variable to accommodate heteroscedasticity (i.e., varying variance). Square root transformation model is used in the first part of the study even though the log transformation is frequently used in the literature [45, 59, 4]. The square root transformation is shown to better stabilize variance. However, Poisson count model is used when investigating multiple counter locations for better comparison and interpretation.

(2) Examine and quantify the nonlinear relationship between weather factors and daily bike volume. It has been noticed that there are nonlinear relationships between weather and bike volume [4, 41, 60, 37]. However, such nonlinearity has not been explicitly modeled and their effects are mostly analyzed using exploratory data analysis. In this thesis, General Additive Mixture approach is adopted to model the nonlinear relationship and investigate its impact on variable selection and model specification.

(3) Include interaction terms in the regression model. There are few papers that have included interaction terms in their modeling. In those that do consider, their attention is only devoted to certain interactions between humidity and temperature, or temperature and wind speed [41]. In this thesis, we extend interaction terms to include both weather and seasonal variables (e.g., precipitation probability and weekend).

3. **Fit an Autoregressive Integrated Moving Average (ARIMA) model to better capture the autocorrelation in the bike count time series**. With the exception of limited number of papers [18, 45], majority of literature assume daily bike ridership count is an independent and identically distributed random variable. In order to develop a better prediction model, ARIMA methodology is fitted to account for possible autocorrelation in the daily bike counts. The accuracy of the resulting model is validated using historical data.

4. **Provide good visualization to show the results, goodness of fit of regressions and their predictive performance**. Due to the inconvenience in interpreting square root transformation model (i.e., no direct explanation in the form of "one percent change in the independent

variable results in certain fixed amount of percent change in the dependent variable", as in the log-linear model), counterfactual simulations are conducted to help visualize the effect of changing one variable while controlling others. In order to test the predict performance, this study use two years of data to fit the model and use the model to predict the third year.

5. **Compare the different impacts of weather and seasonal factors on multiple locations and identify their unique bike travel pattern.** This work extends previous methodology to five different bike locations in Seattle and identifies their unique travel patterns and sensitivities to the same set of factors. These findings will enhance the city of Seattle's ability to determine the actual usage portfolio of bike facilities and help identify locations for new facilities combined with proper geo-spatial information

Chapter 2

LITERATURE REVIEW

Bicycle ridership in cities is useful for practitioners and researchers to understand safety, travel behavior, and development impacts. Therefore the relationship between bicycle volume and various factors, with the goal to build a predictive model based on this relationship, has been of great interest to researchers over the last decade. Among others, weather variables such as temperature and precipitation have long been known to have significant impacts on bike travel demand and travel experience [23], since cyclists are fully exposed to outdoor weather conditions. Following the pioneering work by Hanson et al in 1977 [24], there have been many studies that attempt to explore the relationship between various weather factors and bicycle volume counts (e.g. [20, 17, 43, 45, 55]). With the advent of automatic bike counter that can continuously record the passing bike counts at specific locations, more statistically sophisticated models can be developed to account for more complicated scenarios and offer more explanatory/predictive power. In the following, we summarize existing literature from three perspectives: data source of bike count, variable selection, and modeling methodology. Limitations of existing literature are briefly discussed at the end of this chapter.

2.1 Cycling Ridership Study

There are two types of data sources than are commonly used to explore the relationship between various factors and bike traveling: 1) Travel survey/census data have been either specifically designed to suit the purpose of the study [31] or were broad based travel surveys where data on all travel activity was collected using a successive sample approach over an extended period of time [51]. Bike count data obtained through this type of source is typically used to explain influencing factors such as physical, demographical and socio-economic factors on mode choice [46, 26];

2) Observational travel data that is collected either manually or automatically. The manually collected data is usually collected for a specific purpose of the study [42] and over a short period of time. On the other hand, the automatically recorded data is continuously collected by automatic data collection equipment over a long period of time [20, 45, 41, 59]. One obvious advantage of the automatic data collection systems is that they usually provide a longer time series of data, which will allow modeling of greater variation in weather/temporal parameters [4], whereas special purpose surveys are likely to either be of short duration or rely on respondent's recall of their behavior in the past, which is likely more prone to errors [31]. Supplemented with weather, temporal and other continuous factors, the automatically collected count data is suitable to develop predictive statistical models to forecast bike volume.

2.2 Weather and Seasonal Impacts on Bike Count

A literature review accompanying a recent report by [6] identified eleven primary indicators. These included time of day [56], season [43], population and employment densities [39, 48], land-use mix [48], bicycle facility type [27], traffic volume [40], rain and temperature [43, 46], income [63], and age [27]. This section outlines the key points made in the literature that are relevant to some of most important variables.

Because cyclists are fully exposed in outdoor condition, weather variables play a crucial role in cyclists' decisions to ride. Research has found the variability for counts has a positive association with high temperature and low precipitation [43, 46]. Thomas et al [59] found that temperature caused greatest variation while wind caused the least variation in bicycling demand in Netherlands. Winters et al [65] examined the association of utilitarian cycling with precipitation and temperature in Canadian cities and found that more days of precipitation per year and more days of freezing temperature per year resulted lower utilitarian cycling. According to [41], after other factors controlled for, a 43% to 50% reduction in cycle volumes could be expected when humidity doubled; however, the temperature had a negative effect when it was higher than 28°C and humidity was greater than 60%.

Apart from temperature and precipitation, [41] finds humidity and additional variables includ-

ing the presence of rain in the morning and/or during the previous three hours to be significant too. Gallop and Tse [18] found that rain in the previous three hours can have a lagged effect on cycle counts in the current hour comparable to or greater than rain in the current hour. In the work by Yu et. al [68], the perceived temperature (a combination of air temperature, relative humidity and wind elements), wind speed, visibility and significant weather (a combination of rain, lighting, hail and snow) are considered to model traveling behavior. Other variables that have been examined include hours of sunshine and wind speed [60], and cloud coverage [24].

Moreover, as suggested by [37] and [59], the effects of precipitation and temperature on bicycle volumes are nonlinear. For example, bicycle traffic can decrease in both very cold and very hot weather as noted by [51]. In [51], the author also concluded that the optimal condition for bicycle usage occurs at approximately 25°C with no rainfall, whereas in Phung and Rose [47] the ideal riding temperature is about 28°C, both in the city of Melbourne Australia. Also, interaction effects among different variables have been observed in the literature. For example in [41], the combination of warm and humid weather is found to have large negative effect on bike counts.

Besides weather variables, seasonal factors such as day of week and day light hour are found to have impacts on the bicyclist counts. Day of week has been commonly used in the literature to account for the variation of bike counts during the week [4, 55, 41]. Holiday is also sometimes included to model the abnormal change in bike counts due to holiday events [47]. Other seasonal factors that are considered in the literature include day light hour of the day [4], school in session [55], hour of the day [41] and month of the year [61].

Other comparative studies are also available where bicycle counts are conducted in different cities [45], and different sensitives to weather are examined [53]. As for longitudinal studies, [43] finds increased variability for counts conducted in the later months of the year. [30] conclude that morning peak hours from 6 AM to 9 AM accounts for a consistent 95% of the total bicycle volumes by hourly count data.

2.3 Modeling Methods and Prediction

The vast majority of existing literature use regression-type of models to link the bike counts with various weather and seasonal factors. Standard linear regression with OLS estimation were the most popularly used regression technique [29, 30]. Log transformation in linear model is often used to stabilize the variance [45, 59, 4]. Square root transformation is also a frequent used techniques to accommodate heteroscedasticity in the model. Other regression models used to fit bike count data include Poisson model [43, 41], or negative binomial model [55] where overdispersion is observed.

Bike counts (whether it is hourly, daily or weekly recorded) are often associated with time index. However, there are limited number of papers that focus on developing time series model for bike counts (with the exception of Thomas et. al [59] and Gallop et. al [18]). The work by Gallop et. al [18] incorporates an auto-regressive integrated moving average (ARIMA) analysis. Compared to the normal OLS regression model, ARIMA approach accommodate complex auto-correlation patterns of the error terms. The comparison clearly shows that OLS regression method tends to greatly overestimate the effect of weather variables on bicycle counts.

Researches are also using different perspective to look at bike models. In [41], the authors proposes two modeling approach: the absolute relationship approach and the relative relationship approach. The absolute ridership model, directly relates absolute hourly bicycle volumes to both temporal variables and weather variables, while the relative approach is used to relate deviations from average cycle volumes to corresponding deviations from average weather conditions at the hourly level.

While some studies make no distinction, several have examined the effects of weather on utilitarian and recreational cycling separately [9, 24, 60]. In [45], a model is developed to use deviations in daily weather conditions from average conditions to predict deviations in daily cyclist totals from the average daily total. A summary of important literature surveyed above with their selection of variables and methods are presented in the following Table.

References	Variable							Methond					
	Temp	Precipitation	Wind Speed	Daylight	Humidity	Seasonal	Other	OLS	Logistic	ARIMA	NB	Poisson	Other
Fields-12	Daily Avg.	Weekly Total											No explicit model
Gallop-12	Hourly	drizzle/rain/snow flag	Hourly		Relative humidity	Holiday, Weekend	clearness, fog			✔			
Ahmed-10	Daily Avg.	Daily Total	Daily Avg.	Duration in hrs	Relative humidity	dow		✔					
Nankervis-99						Month	Weather description (i.e., rain, wind, etc).						
Griswold-11							Nearby population and employment density, proximity to downtown/freeway, age, education level, income, etc.	✔					
Helbich-14	Daily Max	Daily Total	Daily Avg.						✔				
Tin Tin-12	Daily/hourly Max.	Daily/hourly total	Daily/hourly Max	Duration in hrs			Facility type	✔					
Hunt-07							Descriptive variables indicating lane use, secured parking, level of experience, etc.		✔				
Jones-10							Length of bicycle network, employment density, population density	✔					
Lewin-11	Daily Max	Rain flag				Weekend		✔					
McCahil-08							logarithmic choice measure, population density, worker density						A new space syntax theory
Moreno-11	Hourly	Hourly, rain presence in prev. 3hrs flag	Hourly		Relative humidity	dow, Month, Year, Hour		✔			✔		
Niemeier-96	High temp. flag	Rain flag				Morning flag, Month category	Location flag					✔	
Nosal-14	Hourly	Rain presence flag (am/pm, prev. 3hrs)			Relative humidity			✔					
Parkin-08							Gender, car ownership, hilliness, off-road routes proportion		✔				
Pinjari-09							Household density, employment density, fraction of commercial land area, demographic factors		✔				
Rose-11	Daily Max	Daily Total				Holiday, school season, dow							Aggregate demand model
Thomas-09	Daily Avg.	Duration of Precip. (in hrs)	Daily Avg.	Duration in hrs				✔					
Schemiedeskamp-16	Daily Max	Daily Max		Duration in hrs		dow, Holiday, UW flag, Day #					✔		
Dunlap-15	Daily Max	Daily Total				dow		✔					
Fagnant-16							Population density, Employment density, bridge, recreational area access, bike trail access				✔	✔	

2.4 Limitations of Current Literature

There are four major limitations in current literature that have been identified and will be address explicitly in this thesis:

1. **Data suitability and model assumption**. Although almost all papers propose a reasonably fitted statistical model, few literature has spent energy in examining model suitability. For example, the justification of taking certain transformation of the dependent variable is sometimes missing. Also, validation of key model assumptions such as homoscedasticity and normality of the residual error are often absent.

2. **Nonlinear relationship**. Despite being mentioned in many studies, the nonlinear relationship between weather factors (such as temperature and precipitation) and bike ridership count has not been formally and explicitly modelled. The squared temperature is often included to capture its nonlinear effect on bicyclist volume [4, 47] and the precipitation volume is often categorize into discrete levels (e.g., light, heavy) for the same reason [47, 60]. A systematic approach to study the nonlinear relationship would be desirable and also helpful for model design across different sites.

3. **Goodness-of-fit**. One limitation present in much of the past literature is that few discuss goodness of fit of their modeling. The model results are mostly interpreted by inspecting coefficients and R squared values. However, a rich body of goodness-of-fit metrics such as out-of-sample fitting and visualization tools are less used in the literature.

4. **Forecasting model**. A model that can better describe and forecast the bicycle count in longitudinal form is necessary to be developed. To that end, however, auto-correlation in the bike count observations is rarely accounted for with the exception of [18] and [45], where a regression model with ARMA error terms is fitted. Forecasting models with well-defined uncertainty quantification would be valuable to have for various transportation planning purpose.

In addition, there is little comparative studies in literature taking into account the impact of different cities/sites, urban form characteristics, topography and cyclist culture on the reactions to different weather conditions. Due to a wide range of factors, cyclists in some cities may exhibit a different response to weather conditions, and it has been suggested that utilitarian cyclist trips are less sensitive to weather than recreational trips [59], but no studies have shown evidence for other factors mentioned above. These topics will be subject of future research and thus out of scope of this thesis.

Chapter 3

METHODOLOGY

In order to discern the relationship between bicycle counts and weather, seasonal factors, we developed a statistical model to predict daily bicycle counts from these other factors. This section describes the data sources such as how it is been collected and processed to prepare for this study. Then exploratory data analysis is presented to show the methods and rationale we used to select variables. Lastly, we introduce the regression models and estimation procedures including ordinary least regression, count model, generalized additive mixture model as well as time series model (ARIMA).

3.1 Data Collection, Processing and Description

3.1.1 Study Location

Data were collected at Seattle's Fremont Bridge and cover a period of more than three years from October 31, 2012 to December 31, 2015. As is shown in Figure 3.1, the Fremont bridge crosses the Lake Washington Ship Canal and links the Fremont and the Queen Ann neighbourhoods. The reasons for picking Fremont bridge as our study location are as follows: 1) A permanent, automatic bike counter is installed at the Fremont bridge, which provides continuous bike counts; 2) the Fremont Bridge represents one of the busiest utilitarian bike facilities in the City of Seattle; 3) The bike counter on Fremont bridge was first installed in October 2012, the earliest site among the nine bike counters in Seattle, and therefore provides a rich dataset; 4) It's relatively close to the University of Washington, and connects the northern part of Seattle to its downtown area; 5) It captures a substantial amount of bicycle traffic due to its status as one of only five facilities that carry bicyclists across the canal separating the northern and southern halves of Seattle.

SDOT has a total of nine automatic bike counters (four of which also count pedestrians) located

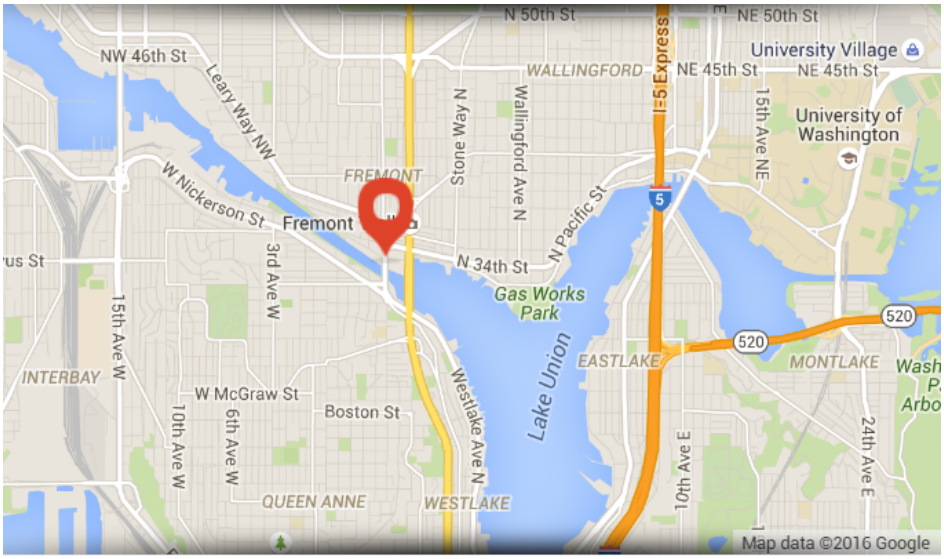


Figure 3.1: Location of Fremont Bridge

on neighborhood greenways, multi-use trails, at the Fremont Bridge and on SW Spokane Street. The counters help create a ridership baseline that can be used to assess future years and make sure the right amount of resources are invested so that the goal of quadrupling ridership by 2030 could be achieved [3]. A multiple location study is included in Chapter 6.

Another important reason to study Fremont Bridge is the recent trend that more and more commuters choose to bike to work. While only 3 percent of downtown Seattle's 200,000 daily commuters now bicycle, the number of bike commuters has increased 18 percent since 2010, according to a survey done for the Downtown Seattle Association [1]. As a major bike site that records bike volumes to and from downtown Seattle, a good understanding of the Fremont Bridge could provide valuable insight on future facility planning to accomodate the growing bicyclist population.

3.1.2 Bike Count Data

Bicycle counts were collected at Fremont bridge continuously by the City of Seattle using an in-sidewalk counter manufactured by EcoCounter (see Figure 3.2). When a bicycle passes over an induction loop embedded in the sidewalk on either side of the Fremont Bridge, the counter registers the bicycle. Data from this equipment has been used in a wide range of studies, and when operating

Figure 3.2: EcoCounter on the Fremont Bridge

properly, the absolute error of these counters has been shown to be below 4% [44]. Bicyclists may legally choose to ride in the roadway instead of the sidewalk, and would thus not be detected by the counter. However, we believe these crossings are rare at this location due to the design of the facility (as is seen in Figure 3.2), which directs bicyclists to enter the sidewalk, and from our own experience riding and observing other riders. The counters collects data 24/7/365, and upload data once a day at 5 am, which is then aggregated into 15 minute intervals by the City of Seattle, and are made available to the public via the City of Seattle's data portal [10, 11].

The bike count data used in this study cover a period of three years spanning from October 31, 2012 to December 31, 2015. The continuous bike count data is aggregated into daily counts. Note that in the literature, there are also studies using hourly bike count. However, the daily bike counts are favored in this study because: 1) it carries less autocorrelation than the hourly data; 2)

it's intuitive and simple to interpret, 3) we believe for commuting cyclists, they are more likely to make decisions of riding based on daily weather, whereas recreational riders are more likely to make decisions on a hourly basis [45].

3.1.3 Weather Data

Weather data are collected by a variety of sources and are aggregated by `Forecast.io`. These data are available through the company's web services API [58]. Historical daily summaries are available for a range of weather variables including several specifically important to our model such as precipitation, daily minimum and maximum temperatures, sunrise, and sunset, etc.

We downloaded and processed these data programmatically using the R programming language along with several add-on packages [21, 64, 13, 34, 50]. Bicycle counts were aggregated by day, and then joined to weather data by date.

3.1.4 Seasonal Data

In addition to the variables collected from the above-mentioned two sources, we were also interested in controlling for holidays and whether or not the nearby University of Washington was in session. These data were collected and coded manually from the National Holiday calendar as well as the University of Washington's academic calendar. Day of week information is also included to explain potential day-to-day bike count variation.

3.2 Descriptive Analysis and Variable Selection

In the following, we selected a subset of variables that we felt best reflected our research questions.

3.2.1 Weather Variable Selection

Weather variables have a huge impact on daily bike counts as cyclists are fully exposed to outdoor conditions. However, there are many different representations and measurements of weather. In

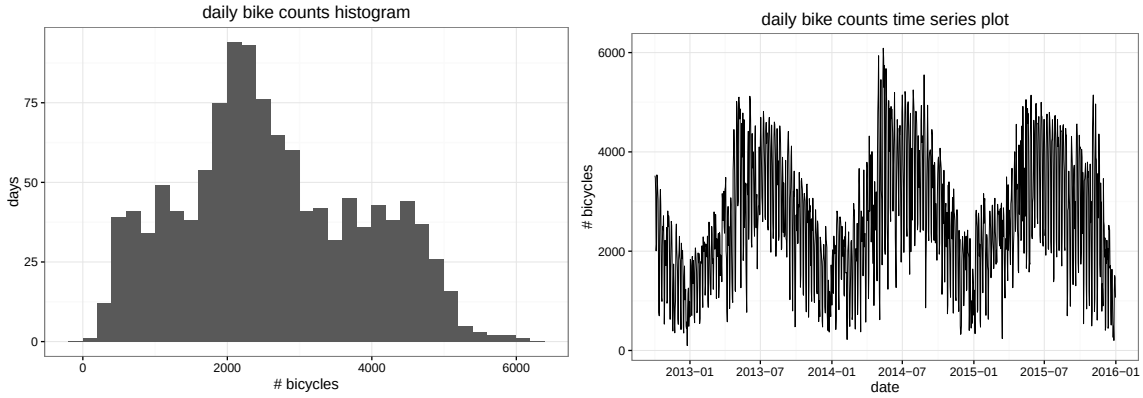


Figure 3.3: Descriptive visualization of bicycle counts dataset: (a) Frequency of observed counts; (b) Timeseries of counts.

this study, weather variables that provide the highest explanatory power for the variation in bike counts are identified. We also include variables that their impact on bike counts are very intuitive.

Literature has unanimously agreed that temperature and precipitation are the two main weather determinants for bike ridership [45, 61]. In general, cycle counts increases with temperature. However, with daily-aggregated weather information, decision needs to be made regarding which form of the temperature variable should be used in the model, e.g., the daily maximum temperature or the average temperature. In the literature, arguments have been made for both cases. For example, in the work by Tin Tin et. al [61], the maximum temperature is used because people are most sensitive to the extreme, while in Thomas et. al [59], the mean temperature is used because the authors believe that many cyclists make their trips in the morning, during which the temperature lies closer to the mean. In this study, daily maximum temperature `TempMax`, measured in Fahrenheit, was chosen to represent temperature in part to retain simplicity in the model, in part because there is relatively little daily temperature variation in Seattle due to the moderating effect of large water bodies, and in part because maximum temperature better reflects the conditions during daylight hours when most bicycle trips would occur. This simplification may not be warranted for other locations that experience greater temperature variation than Seattle.

Precipitation, or rainfall, also has a significant impact on the cycling counts. However, sim-

ilar to temperature, there are multiple choices of the precipitation variable in the original dataset that could be used in the model, such as the precipitation amount [4], duration of the precipitation [59], rain presence [41], etc. In this study, the precipitation probability `PrecipProb`, which measures the probability of precipitation for the day, was chosen over other alternatives based on the assumption that bicyclists usually make travel decisions based on the likelihood of raining *ahead* of the day, while variables such as maximum precipitation or precipitation duration in the `Forecast.io` report only represent the fact *after* the rainfall have occurred. As in the case of temperature, this simplification would be less justifiable in locations that experience greater daily variation in precipitation or in locations that have a predictable pattern of precipitation during certain hours.

To support our choice on `TempMax` and `PrecipProb`, exploratory data analysis for continous varibale as correlation scatter matrix is depicted as in Figure 3.4. To avoid collinearity it is reasonable to keep only `TempMax` because it has the highest correlation with bike counts. Similarly, in the scatter matrix Figure 3.5, it is noted that correlation between the probability of precipitation and bike count is -0.452, strongest of all precipitation-related variables.

Furthermore, we fit models with different forms of the temperature and precipitation variables, with all other variables held the same. For temperature, we compare the `TempMax`, `TempMin`, `TempAvg`, as well as the apparent temperature `ATempMax`, `ATempMin` and `ATempAvg`. For precipitation, we consider `PrecipProb`, `PrecipIntensity` and `PrecipIntensityMax`. R^2 values are used to compare the model performance, which indicates the explanatory power of the different variables. The results are summarised in Table 3.1. It is seen that the `TempMax` and `PrecipProb` have the highest R^2 value and therefore the highest explanatory power. These results confirm our choice of `TempMax` and `PrecipProb`.

Besides temperature and precipitation, other weather variables have been identified in the literature to have different levels of influence on cycle counts. In the work of Moreno and Nosal [41], humidity is found to have a negative relationship with cycling. The factor of wind speed is shown to be significant in [60], while in the tests by [41] wind speed had no impact on the results. Other variables that have been examined include sunshine [60], cloud coverage [24],

Temp. Variable	TempMax	TempMin	TempAvg	ATempMax	ATempMin	ATempAvg
R^2	0.866	0.843	0.858	0.864	0.844	0.857
Precip. Variable	PrecipProb	PrecipIntensity	PrecipIntensityMax			
R^2	0.866	0.860	0.858			

Table 3.1: Goodness-of-fit results of models with different forms of temperature and precipitation variables, all other variables controlled.

dew point [54, 45] and visibility [59]. Considering their relatively less significant impact on cyclist counts and for the purpose of simplicity, variables `cloudCover`, `dewPoint`, `humidity`, `moonPhase`, `visibility`, `windBearing`, `windSpeed` are excluded in further analysis.

The categorical variable `Weather` is investigated. It represents the general weather classification, such as clear-day, cloudy, foggy, rainy, windy, etc. The assumption behind this is that bicyclists are likely to make their travel decision based on the general weather description. The `Weather` variable serves as the summary of weather conditions and could potentially account for nonlinear relationship and reduce model complexity.

3.2.2 Seasonal factors

Seasonal factors are also found to have significant impact on bike counts in the literature. Such impact may be caused by weather, holidays, working schedules and school calendars, etc.

Daylight hours `Daylight` (defined as sunset time $-$ sunrise time) was selected to represent seasonality. One justification for choosing daylight hours over the calendar-based categorization of season is in part because Seattles Pacific Maritime Climate differs substantially from traditional notions of four seasons. Daylight hours also is measured as a continuous value at a finer temporal resolution of one day. Finally, daylight hours adjusts according to latitude, which may make this model estimation procedure and specification more transferable to other sites in the future, perhaps by interacting latitude with daylight hours. For the sake of comparison, the traditional `Season` variable is also included, which takes values in Spring, Summer, Autumn and Winter as categorical variable.

University of Washington in-session status `UW` was selected to represent seasonality associ-

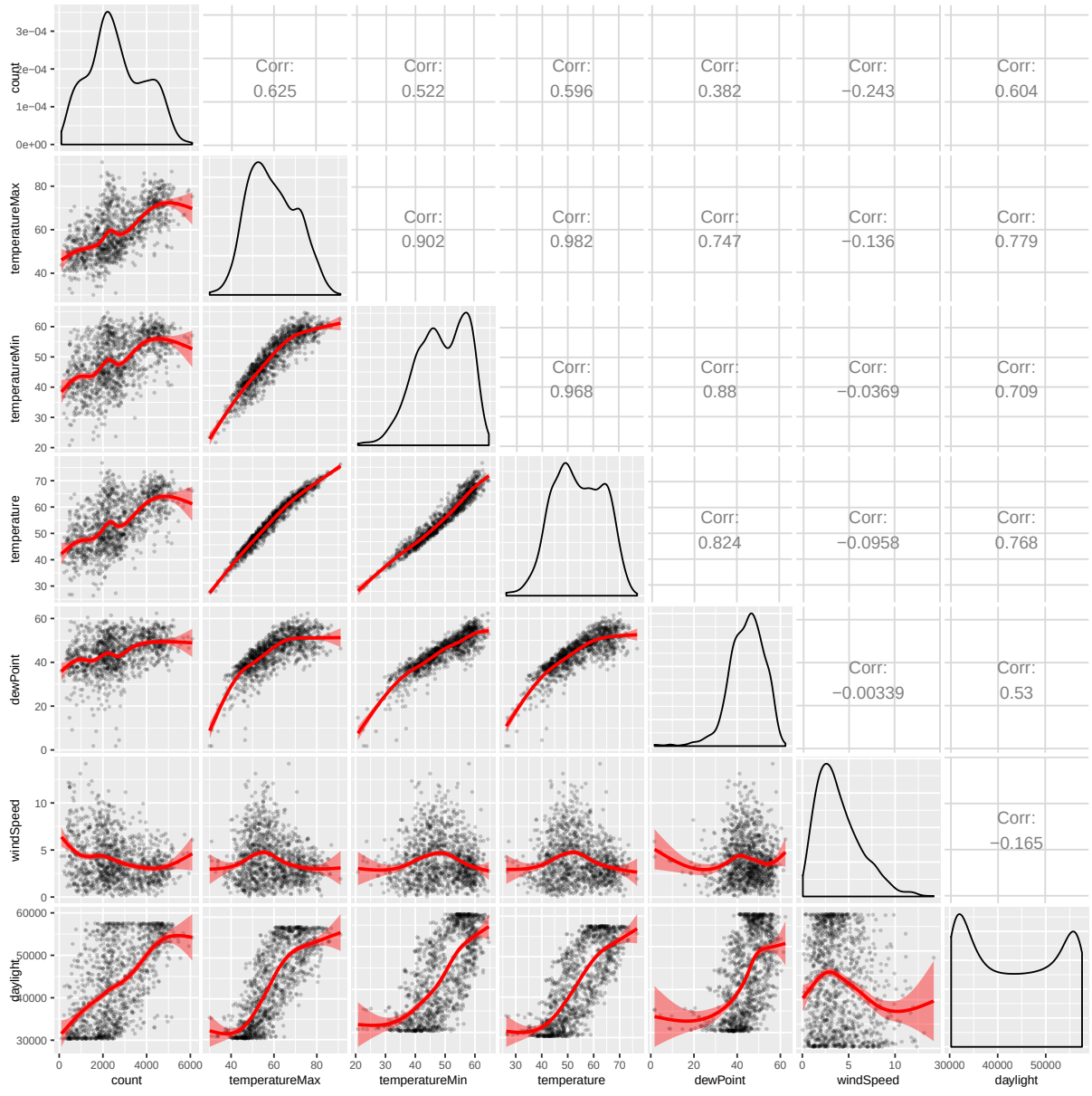


Figure 3.4: Scatter plot Matrix of covariates in the Group of Temperature.

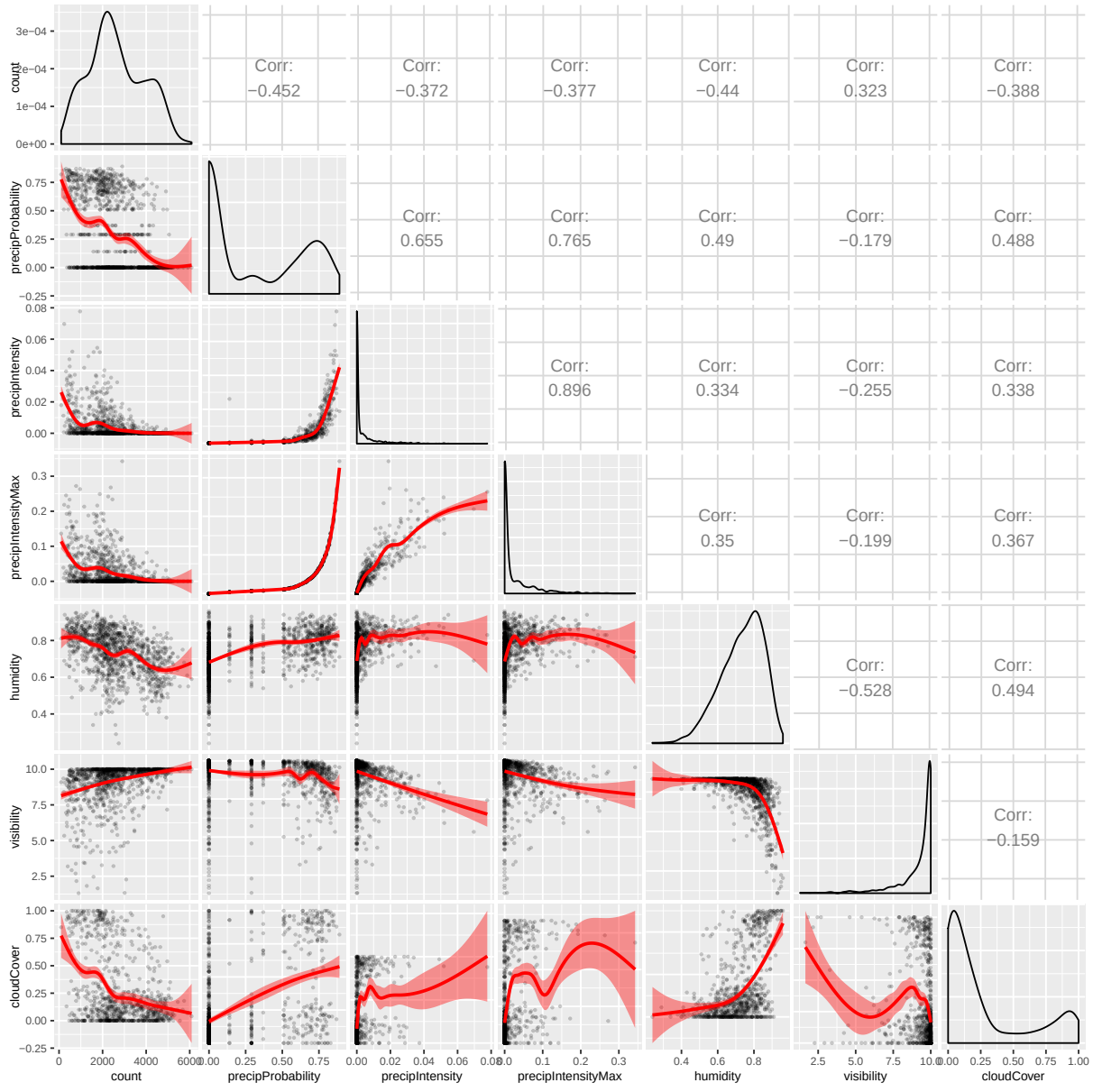


Figure 3.5: Scatterplot Matrix of covariates in the Group of Precipitation.

ated with the UW. We deemed the University of Washington variable important in part because of the Fremont Bridge's proximity and connection via the Burke Gilman Trail to the University of Washington. We also felt that this variable was a suitable proxy for the school season, which more broadly captures whether or not other local schools are in session. The academic calendars of the various local schools do not align perfectly, however they still overlap substantially with the University of Washington, which is itself the largest educational institution in the region. Moreover, since school buses serve as the predominant transportation mode in elementary/middle/high schools, it is fair to assume most student cycling traffic comes from the UW.

Inclusion of the holiday variable `holiday` was an attempt to account for some low outlier counts. Upon inspection of the dataset, Christmas and Thanksgiving in particular had very low counts of bicycles relative to the days preceding and following. Relatedly, but not accounted for by any variable in our model, are some of the high outlier counts. Upon inspection, some of the highest counts were observed on National Bike to Work Day and on the day of the Fremont Solstice Parade, which typically draws large numbers of bicyclists as participants and spectators. The omission of such a variable is justified based on the few occurrences of high outlier counts, and our desire for this model to only include variables that could be collected or straightforwardly adapted to other locations.

Day of the week `dow` was added due to its presence in the literature, as well as an apparent weekly pattern is revealed visually by zooming into the time series plot (Figure 3.3). These data were coded as a set of categorical variables. In addition, the weekend or not flag `Weekend` is included since we are interested in understanding the traveling behaviour of bicyclists at Fremont Bridge, mostly commuters. It is expected to have big variation in bike counts between weekends and weekdays but not so much between individual day of the week.

The final variable, the time index `day`, was included so that we could test for a linear trend in bicycling volumes. We created this variable by sequentially numbering (1–1157) the observed counts by day during the study period.

Box plots for categorical factors against daily bike count are depicted in Figure 3.6. Visually, we are able to observe the difference in their means, suggesting these variables have explanatory

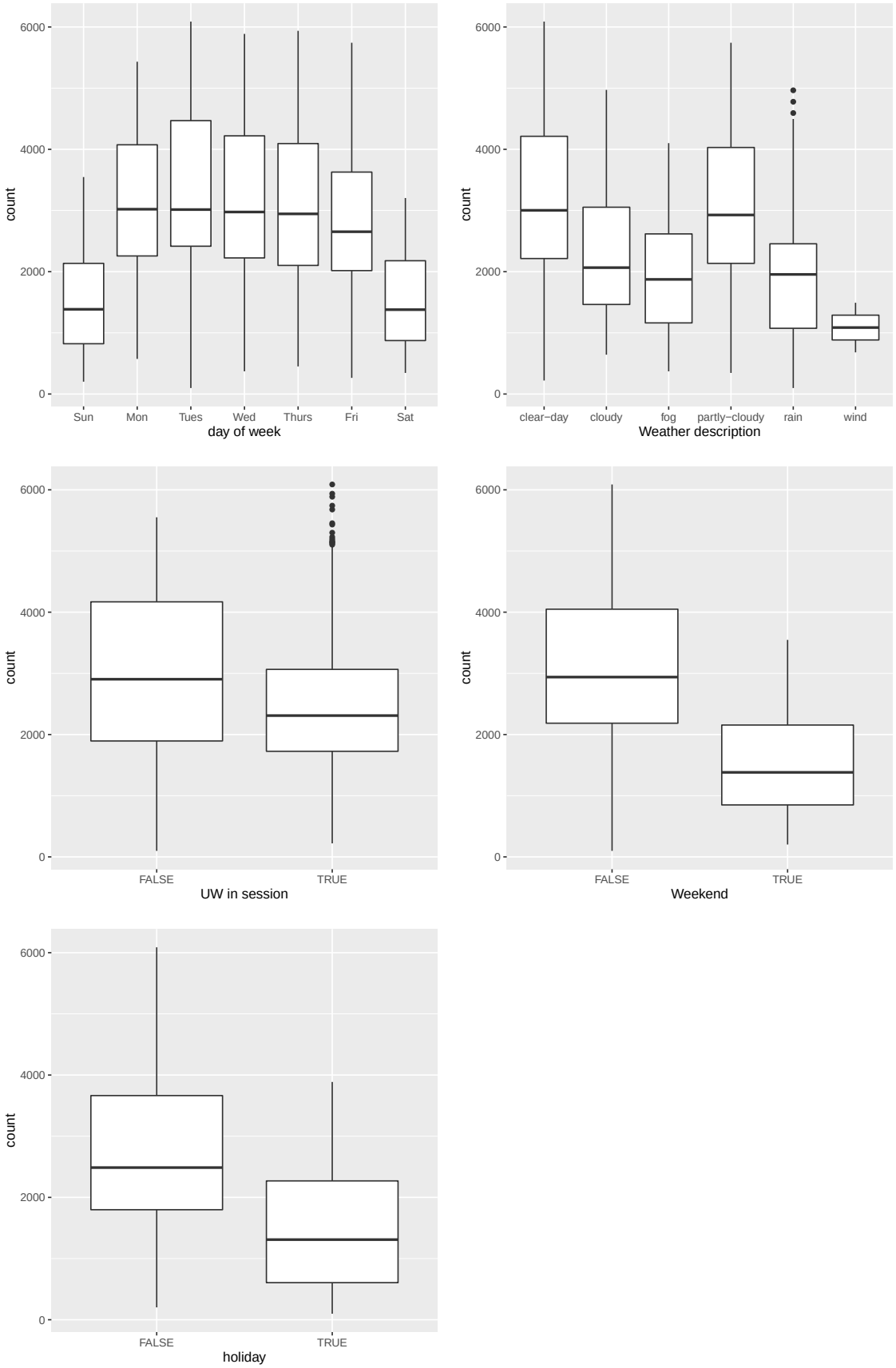


Figure 3.6: Count box plots of `dow`, `Weather`, `UW`, `Weekend`, `Holiday`.

Table 3.2: Candidate explanatory variables

Variable	Description
`Count*`	Number of bicycles per day
`TempMax`	The maximum temperature for the day
`PrecipProb`	Probability of precipitation for a given day
`Daylight`	Time from dawn to dusk in hour
`dow`	Day of the week categorical variable
`Holiday`	The day was a holiday (TRUE/FALSE)
`Weekend`	The day is a weekend (TRUE/FALSE)
`UW`	The University of Washington was in session (TRUE/FALSE)
`Season`	Season indicator spring, summer, autumn, and winter
`Weather`	General weather classification, such as clear-day, cloudy, foggy, rainy, windy, partly-cloudy-day, partly-cloudy-night (7 levels)
`day`	Time index (sequentially numbered) during study period

* Dependent variable

power for variation in bike count. Formal analysis on their impacts is presented in later Chapter 4.

3.2.3 Summary of this section

Combining above analysis, a set of explanatory variables is extracted from the original dataset. Their names and descriptions are summarized in Table 3.2. It represents the group of variables that we are interested in examining and will be further investigated in the latter sections.

3.3 Modeling Approach

There are multiple approaches to model the relationship between weather and seasonal variables and daily cyclist volumes. The simple linear regression model is first reviewed, and compared with some of its most frequently used variations, as well as the count model (i.e., Poisson model). The choice of preferred model is then made based on purpose of analysis and model residual error comparison. To account for the nonlinear relationship between weather and bike counts, the generalized additive model (GAM) is adopted and briefly reviewed in this section. Lastly, an autoregressive integrated moving average (ARIMA) model is proposed for prediction of daily bike counts, and the autocorrelation is taken into account.

3.3.1 Regression Model

Linear Model

The ordinary least squares-based linear regression is one of the most widely used techniques in Statistics to model the relationship between dependent and explanatory variables. The goal is to minimize the differences between the observed responses and the predicted response given by the linear approximation of the data. The basic linear regression model assumes the following structure:

$$Y_i = \alpha + X_i^T \beta + \varepsilon_i, \tag{3.1}$$

where Y_i is the response variable (or observations, dependent variable), and X_i is the predictors (or regressors, independent variables), α and β are the unknown parameters to estimated, and ε_i is the

unobserved scalar random variables (errors) which accounts for the discrepancy between the actual observed responses Y_i and the predicted responses $\alpha + X_i^T \beta$.

The OLS technique offers a mathematically convenient tool to estimate the linear regression model parameters α and β. There are many available software packages to provide solutions to the linear problem (3.1), i.e., the `lm` function in `R`. However, to properly apply the OLS estimators, certain assumptions need to be checked beforehand, such as: 1) No perfect linear dependence, 2) Strict Exogeneity ($E[\varepsilon|X] = 0$), 3) Homoscedasticity ($E[\varepsilon^2|X] = \sigma^2$), and 4) Normality (the error term has a normal distribution).

The goodness-of-fit of the considered model is often evaluated with the R^2 (R squared, or the coefficient of determination) and adjusted R^2. The R^2 measures the percentage of the response variable variation that is explained by a linear model, or equivalently

$$R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}}. \tag{3.2}$$

The adjusted R^2 adds a correction for the number of estimated parameters to guard against over-fitting. Other important goodness-of-fit criterion that are used in this study are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The smaller AIC and BIC value is, a better fit of model we have.

Note that the dependent variable in linear regression model is usually continuous-valued, which is not the case with bike counts. Nevertheless, recall that in Figure 3.3 of Section 3.2, the distribution of daily bike counts follows an approximate Gaussian distribution. Moreover, the daily bike count data we collected over the past three years is very large (the median is over 2000) and it could be treated as if it was continuous-valued. Therefore the linear regression model is considered in this thesis.

Count Models

Since the daily bike count data is non-negative and discrete-valued, a natural model choice to fit the bike count data would be a count model, such as Poisson model [43] or the negative binomial model [55] if over-dispersion is observed in the data. Poisson model assumes the dependent vari-

Table 3.3: Four types of models

Method	Regression equation
Standard linear regression	$Y_i = \alpha + X_i^T \beta + \varepsilon_i$
Log-linear model	$\log Y_i = \alpha + X_i^T \beta + \varepsilon_i$
Square root transformation	$\sqrt{Y_i} = \alpha + X_i^T \beta + \varepsilon_i$
Poisson model	$\log E(Y_i \mid X_i) = \alpha + X_i^T \beta$

able has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model and is a special case of the generalized linear models (GLM) with the logarithm as the (canonical) link function, and the Poisson distribution function as the assumed probability distribution of the response. The mathematical form of Poisson model is as follows:

$$\log E(Y_i \mid X_i) = \alpha + X_i^T \beta . \tag{3.3}$$

The set of parameters in the Poisson model (3.3) can be estimated using maximum likelihood method. Readily available function `glm` can be found in `R` to efficiently fit the Poisson model. Poisson model assumes a Poisson distribution of the dependent variable, which is often violated by the real data, i.e., over-dispersion of the variance. In this case, other generalized linear model such as negative binomial model may function better.

Note that besides the simplicity of Poisson model, another advantage with Poisson model is its ease of interpretation. With one unit increase in one of the explanatory variable, a percentage change in the dependent variable can be straightforwardly calculated.

Choice of Model

In this section, we seek to find the proper model that fits our research question. Both the linear model and count model are examined and compared to each other in terms of residual error and goodness-of-fit results. For linear model, different transformations on the dependent variable are investigated to satisfy the OLS requirements. The Poisson regression model serves as a representative of the count model.

It is known that certain model requirements need to be met when OLS-based linear regression model is applied , i.e., homoscedasticity, normality assumption, etc. When some of the key assumptions are violated, certain type of transformation of the dependent variable can be used in order to reduce skewness or other distributional features that complicate analysis. Common transformations include logarithm and square root transformation. To figure out the proper model, we first fix the explanatory variables X to be `TempMax`, `Holiday`, `PrecipProb`, `Weekend`, and `Daylight`. Then the following choices of model are considered: standard linear regression, log-linear model (with logarithm transformation), square root transformation model and the Poisson model. Their corresponding regression equations are described in Table 3.3.

The following residual plots are used to evaluate the suitability of different types of models:

1. **Residual vs. Fitted plot:** Ideally, the plot of the residuals against the fitted values should show no discernible pattern. If a pattern is observed, there may be heteroscedasticity in the residual error. That is, the variance of the residuals may not be constant. It is shown in Figure 3.7 that the standard linear regression model with original bike count as dependent variable has a clear varying variance in its residual: the residual becomes larger when the bike counts to be predicted become larger. Among the four models we tested, the square root transformation model results in the best residual plot, which remains constant for all fitted values.

2. **Scale-Location plot:** It depicts the square root of the absolute values of the residuals against the fitted values, with a Lowess curve helpfully overlaid. The scale-location plot is often

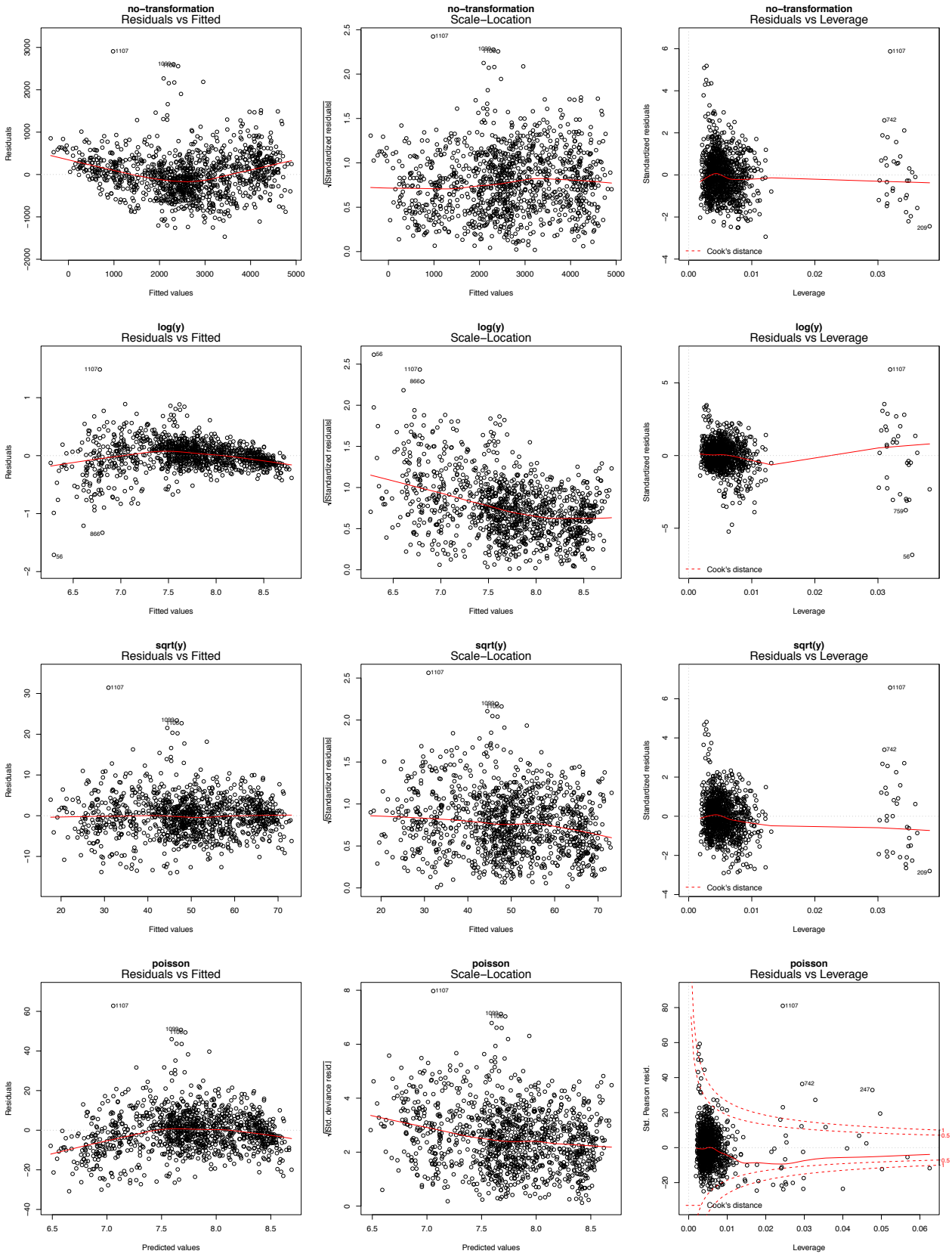


Figure 3.7: Residual plots for different types of models

Table 3.4: Goodness-of-fit results of models

Model Name	Standard Linear Model	Log-linear model	Square root transformation Model	Poisson Model
R^2	0.84	0.82	0.86	0.85
RMSE	501.46	541.14	477.48	495.28

used to check if the data possesses homoscedasticity. Ideally if the modeling data has homoscedasticity, the Lowess curve is expected to be flat, not sloped, and the square root of residuals should be approximately evenly distributed along the Lowess curve. Based on this criterion, it is shown in Figure 3.7 that the square root transformation model has the least heteroscedasticity since its Lowess curve is the flattest and the residual didn't show a discernible pattern.

3. **Residual vs. Leverage plot:** This plot depicts the standardized residuals against the leverage for each point in the data series. The Cook's Distance is also shown in the plot. This plot is mainly used to identify extreme points and possible outliers in the data series that could shift the regression line significantly. The further out the point is on the X or Y axis, the more leverage or standardized residual the point has. More details on the Residual vs. Leverage plot could be found in [22]. From Figure 3.7 it can be seen that there are a few points in the data set that have large leverage on the regression lines (indicated by points at the far right side of the plot). The Poisson model has relatively larger residuals with quite a few points with big Cook's distance. For each model, there is one point has big positive error (1107), suggesting it could be an outlier.

Residual plots suggest the square root transformation model to be the best fit. We also applied the Box-Cox test [8] to find out the optimal transformation parameter λ such that the transformed dependent variable (defined as $T(Y) = (Y^{\lambda} - 1)/\lambda$) follows an approximately normal distribution. As is shown in Figure 3.8, the optimal transformation parameter lies between 0 and 1, and is approximately 1/2, which confirms our previous conjecture.

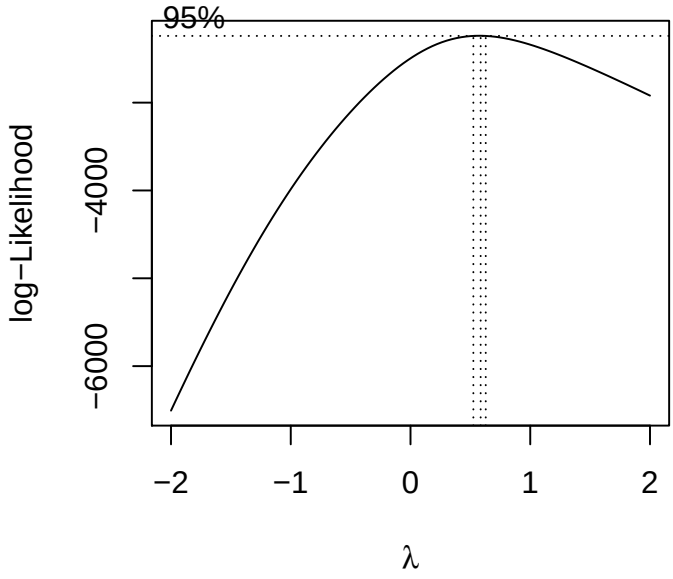


Figure 3.8: Box-Cox test

Furthermore, the goodness-of-fit results of four candidate models are compared and summarized in Table 3.4. The square root transformation model has the highest R^2 and smallest mean squared error, suggesting a better fit than the other three models.

In summary, the square root transformation model is preferred to fit the bike count data at Fremont Bridge: after applying squared root transformation to the dependent variable, the resulting model residuals have an approximate constant variance; error terms follow an approximate Gaussian distribution; the homoscedasticity assumption is satisfied. However, in the multiple location study (see Chapter 6), Poisson regression model is chosen because it better fits the data distribution of other bike sites.

3.3.2 Generalized Additive Mixture (GAM) Model

To account for the nonlinear relationship between weather and bike counts, the generalized additive model (GAM) will be adopted in this work. GAM is also a generalized linear model in which the linear response variable depends on unknown smooth functions of the independent variables. The goal is to provide characterization about these smooth functions. GAM was first proposed in [25] to blend the properties of generalized linear models with additive models.

Following a similar approach with the GLMs, an exponential family distribution (could also be normal, Poisson, negative binomial, etc) is specified for Y along with a link function g relating the expected value of Y to the predictors X_i such as

$$g(E[Y]) = \beta_0 + s_1(X_1) + s_2(X_2) + \ldots + s_m(X_m)$$

The function $s_i(X_i)$ may be specified parametric functions (e.g., polynomial) or may be specified non-parametrically, or semi-parametrically, simply as smooth functions, to be estimated by non-parametric means. The nonparametric GAM provides a very general modular estimation method capable of using a wide variety of smoothing methods to estimate the $s(X)$. The advantage of non-parametric models is that they are easy and efficient to fit, while the disadvantage is the inability to control the complexity of the model (degree of smoothness of $s(X)$), which often gives rise to problems with interpretation. Overall, a well calibrated GAM is likely to perform better than nearly

any other model type, if the dataset is large enough and its behavior is complex enough. However, it could also have problems of overfitting as the number of smoothing parameters increases.

In this work, we use GAM model to explore the nonlinear relationship between the dependent variable bike counts and weather factors (i.e., temperature and precipitation). It is noticed in other studies [47] that the temperature has a positive effect on bike ridership when it is mild, and a negative impact when it gets extremely hot. Also, the temperature squared is often used in the literature [51] to account for the nonlinearity without good explanation. The GAM provides a good starting point to investigate such nonlinear relationship since it explicitly models the dependent variable as a smooth function of explanatory variables. The non-parametric smooth function provides valuable insight on how to include nonlinear terms in the recommended model. We will use the `gam` function in `R` to fit the model.

3.3.3 Autoregressive Integrated Moving Average (ARIMA) Model

An ARIMA-type model is proposed for the prediction of daily bike counts, in which the autocorrelation is taken into account. The ARIMA method usually consists of three major components: differencing, autoregressive model and the moving average mode. The differencing step is used to convert the time series to be stationary, which means there is no predictable patterns in the time series in the long run. The Autocorrelation Function (ACF) plot is also useful to identify non-stationary series. The ACF plot depicts the autocorrelation between the Y_t and its lagged values Y_{t-k} for different values of k. By definition of stationarity, for a stationary time series, its ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly. Other two popular statistical stationarity tests are the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, both available in `R`.

In a pure autoregressive model, the dependent variable Y consists only of lagged values of itself. In other words, we forecast Y using a linear combination of past values of the variable. A moving average model, however, uses the past forecast errors in a regression-like model. It means that each value of Y_t can be thought of as a weighted moving average of the past few forecast errors.

By combining the above three components: differencing, autoregressive model and the moving

average model, we have the non-seasonal ARIMA model. The mathematical expression for the full model is as follows:

$$y'_t = c + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \ldots + \phi_p y'_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} + e_t, \qquad (3.4)$$

where y'_t is the first-differenced time series, and e_t is the white noise error. Equation (3.4) is referred to as the ARIMA(p,d,q) model, where p, d, q is the order for autoregressive, differencing and moving average model, respectively. To choose an appropriate order (p,d,q) for ARIMA, one can use a combination approach of automated procedure `auto.arima` in `R` and direct observations of ACF and PACF plots. Recall that the PACF measures the relationship between y_t and y_{t-k} after removing the effects of other time lags in between: $1,2,3,\ldots,k-1$. The best model is chosen according to the smallest AICc [28]. Note that it is also not common to have an ARIMA model of order more than 3.

Last but not least, it is also possible to include exogenous regressors in the ARIMA model. The mathematical expression for ARIMA with regressors is as follows:

$$Y_t = \beta_t X_t + n_t$$

where X_t is the vector of regressors and n_t is the an ARIMA(p,d,q) model. In other words, the ARIMA(p,d,q) model is fitted to the errors of the regression of y on X (i.e., the series $Y_t - \beta_t X_t$). More details on ARIMA and its implementation could be found in [5].

Chapter 4

WEATHER AND SEASONAL IMPACTS

This chapter discusses impacts of different factors on cycling count. We will use the selected variables and linear regression model discussed in previous chapter to fit bike count data. Predictive models will be discussed in next chapter.

In the following, we will examine non-linearity, discontinuity as well as interaction terms. A Generalized Additive Mixture Model (GAM) is first used to model the nonlinear relationship between bike counts with temperature and precipitation. In particular, Model 0 (base model) and Model1 are developed to address nonlinearity. Model 2, Model 3 and Model 4 include variables that represent measurements of weather and seasonality on different scale (from continuous to categorical or to discrete characteristics). Model5 tests interaction relationship between weather and season. In addition, Model6 quantifies the increase of cycling over time while controlling all the weather and seasonal factors. Model7 ties everything together and represents the 'Master model'. A summary of all model specifications considered presented in Table 4.1.

4.1 Generalized Additive Mixture (GAM) Results

It has been documented in literature that temperature and precipitation have nonlinear impacts on bike ridership [51, 47]. In order to explicitly model nonlinearity, a generalized additive mixture (GAM) model is fitted. The general GAM methodology is described in Section 3.3.2. In this study, we consider a GAM model of the following form:

$$\sqrt{E[\text{Count}_t]} = \beta_0 + \beta_1 s_1(\text{TempMax}_t) + \beta_2 \text{Holiday}_t + \beta_3 s_2(\text{PrecipProb}_t) + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t. \tag{4.1}$$

where s_1 and s_2 are two nonlinear smooth functions of `TempMax` and `PrecipProb`, respectively. A square root link function is used to connect the expected bike counts with the explanatory vari-

Table 4.1: Model specifications

Name	Model Specification with Major Interest
Model0	Base model : Include Non-linearity of `TempMax`
	$\sqrt{count} \sim$ **TempMax**$^2 + TempMax + holiday + PrecipProb + Weekend + Daylight + UW$
Model1	Explore Non-linearity of `PrecipProb`
	$\sqrt{count} \sim$ TempMax$^2 + TempMax + holiday +$ **PrecipProb + ppp** $+ Weekend + Daylight + UW$
Model2	Explore the impact of different weather type : `Weather`
	$\sqrt{count} \sim$ TempMax$^2 + TempMax + holiday + PrecipProb + Weekend + Daylight + UW +$ **Weather**
Model3	Explore the impact of different season : `Season`
	$\sqrt{count} \sim$ TempMax$^2 + TempMax + holiday + PrecipProb + Weekend + UW +$ **Season**
Model4	Explore Day of Week : `dow`
	$\sqrt{count} \sim$ TempMax$^2 + TempMax + holiday + PrecipProb + Daylight + UW +$ **dow**
Model5	Explore the interaction betweeb `PrecipProb` and `Weekend`
	$\sqrt{count} \sim$ TempMax$^2 + TempMax + holiday +$ **PrecipProb** $*$ **Weekend** $+ Daylight + UW$
Model6	Quantify trend by time index variable `day`
	$\sqrt{count} \sim$ TempMax$^2 + TempMax + holiday + PrecipProb + Weekend + Daylight + UW +$ **day**
Model7	Master model: ties together all elements explored above
	$\sqrt{count} \sim$ TempMax$^2 + TempMax + holiday + PrecipProb + ppp + dow + Daylight + UW + Weather + day$

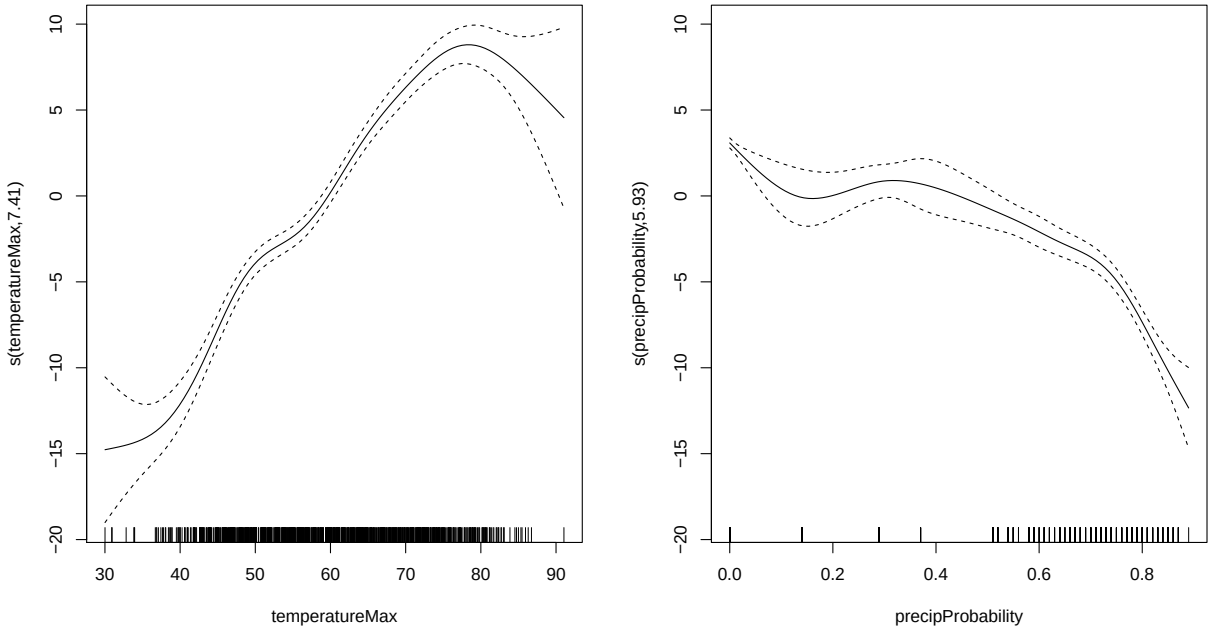


Figure 4.1: Smooth functions of `TempMax` and `PrecipProb`

ables. The `gam` function of `mgcv` package in `R` is used to fit the model. By default, s_1 and s_2 will take the form of the built-in nonparametric smooth splines. The smoothing parameter is chosen automatically using cross-validation. Therefore, the `gam` solution serves as the best approximation of the true nonlinear relationship.

4.2 Weather Impact

We are interested to see the form of the smooth function s_1 and s_2 of temperature and precipitation probability, which characterize their nonlinear relationship with daily bike counts. As is seen in Figure 4.1, the smooth function of `TempMax` clearly follows an nonlinear relationship. The smooth function s_1 provides an accurate characterization of the nonlinear relationship between `TempMax` and bike counts. Approximately, it suggests a quadratic term $\texttt{TempMax}^2$ could be included in the base model to capture the observed nonlinear relationship. Furthermore, GAM helps one to determine important turning points. when he maximum daily temperature is in the range of 40°F and $75^{\circ}F$, the bike counts monotonically increases as the temperature rises. However, when the

max temperature is above 75°F, the increase of bike counts according to max temperature slow down. It is consistent with finding that is reported in other studies [47, 51]. Such implications could be used to generate a categorical temperature variable (i.e., low, medium and high) to reduce complexity in modeling.

Similarly, the GAM model suggests a nonlinear relationship between `PrecipProb` and bike counts which is fully characterized by the smooth function s_2 as is in Figure 4.1. It could be roughly approximated by a piecewise linear function: when the precipitation probability is smaller than 0.75, there is a relatively small negative linear relationship between `PrecipProb` and bike counts; whereas when `PrecipProb` becomes higher than 0.75, the bike count will experience the steeper decreasing rate. This finding is interesting in the sense that bikers at the Fremont Bridge tend to ignore rainfall when make their travel decisions, unless the raining probability is relatively high (>0.75). It also suggests including a piecewise linear function of the `PrecipProb` in the model could be used to account for the nonlinear relationship.

Based on the findings from the GAM fit, we consider the following:

Model0

$$\begin{aligned}\sqrt{\text{Count}_t} = \beta_0 + \beta_1 \text{TempMax}_t^2 + \beta_2 \text{TempMax}_t + \beta_3 \text{Holiday}_t + \beta_4 \text{PrecipProb}_t \\ + \beta_5 \text{Weekend}_t + \beta_6 \text{Daylight}_t + \beta_7 \text{UW}_t + \varepsilon. \end{aligned} \tag{4.2}$$

Model0 includes the $\texttt{TempMax}^2$ to account for the nonlinearity in temperature on bike counts. The fitted parameter estimates, goodness-of-fit metrics are summarized in Table 4.2. Each of the coefficients in Model0 is statistically significant at the $p < 0.01$ level. The adjusted R^2 of Model0 is 0.866, which means approximately 87% of the bike count variance could be explained by the fitted responses of Model0.

In order to provide results that are more readily interpretable by non-statisticians, counterfactual simulations are used to isolate individual terms from the model that correspond to our research questions. In so doing, we simulated various quantities of interest including point estimates and confidence intervals, and then plotted them for visual inspection. More specifically, the counterfactual simulation is conducted as follows: first, the model is fitted and point estimate $\hat{\beta}$ and variance

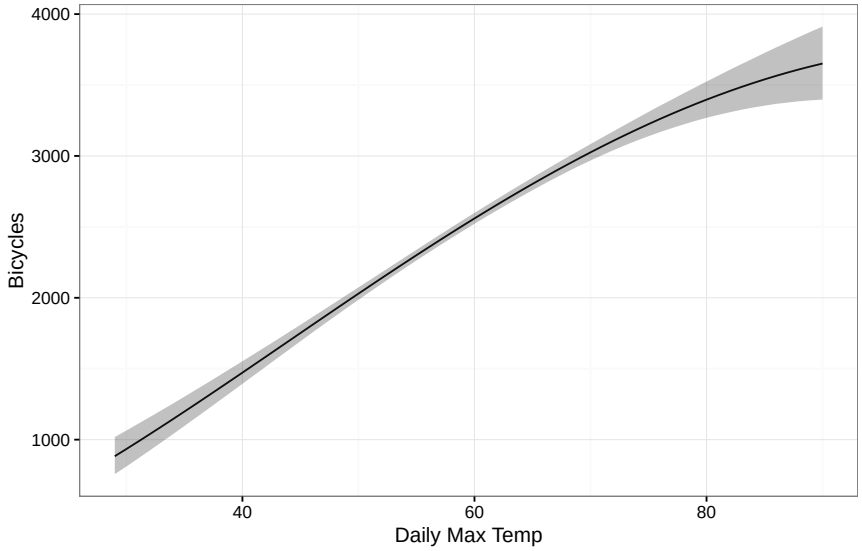


Figure 4.2: Effect of `TempMax` on bicycle counts, all other factors held constant, with shaded 95% confidence interval.

$V(\hat{\beta})$ for the coefficients are obtained; secondly, a counterfactual of the independent variable X_c is chosen; thirdly, certain coefficient $\tilde{\beta}$ is drawn from the normal distribution $\mathcal{N}(\hat{\beta}, V(\hat{\beta}))$ and the corresponding expected response variable $E(Y_i|X_c)$ is calculated; the last step is repeated many times to obtain a vector of expected response variables, and its mean and confidence intervals are then calculated.

Figure 4.2 shows that the temperature variable has a clear positive association with increased number of bicyclists. In this case, the counterfactual X_c is created by varying `TempMax` while holding all other variables constant (i.e., set to their sample means). The nonlinear relationship between temperature and bike counts is accounted for by the squared temperature max term. Such effect is captured for Model0 in Figure 4.2 where the bike counts start leveling off at very high temperatures (higher than 80°F). Because of the relative simplicity, Model0 will be served as the 'base model' in the following discussion.

Next we take the nonlinearity in `PrecipProb` into account by assuming a piecewise linear structure, as suggested by the GAM fit. The alternative model is therefore given by the following:

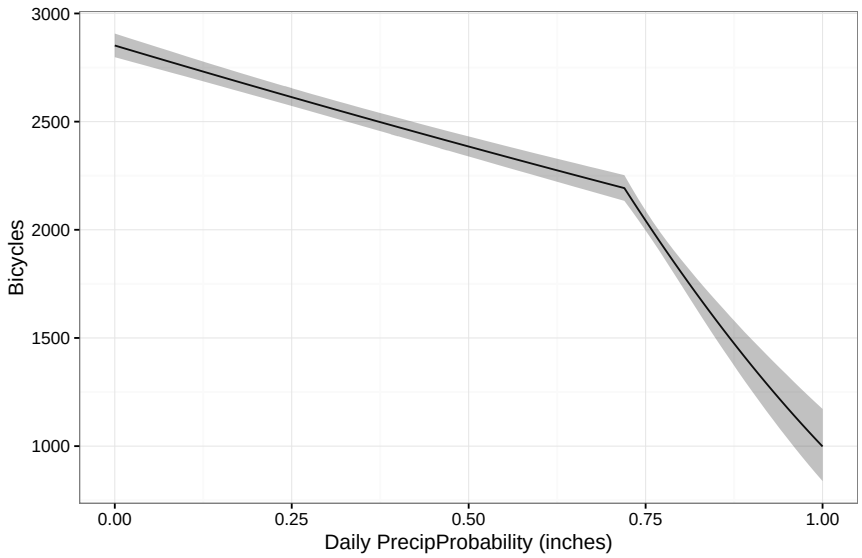


Figure 4.3: Effect of `PrecipProb` on bicycle counts, all other factors held constant, with shaded 95% confidence interval.

Model1

$$\sqrt{\text{Count}_t} = \beta_0 + \beta_1 \text{TempMax}_t^2 + \beta_2 \text{TempMax}_t + \beta_3 \text{Holiday}_t + \beta_4 \text{PrecipProb}_t + \beta_5 \text{ppp}_t + \beta_5 \text{Weekend}_t + \beta_6 \text{Daylight}_t + \beta_7 \text{UW}_t + \varepsilon, \tag{4.3}$$

where `ppp` is the auxilary variable defined as:

$$\texttt{ppp} = \begin{cases} 0; & \texttt{PrecipProb} \leq 0.72 \\ \texttt{PrecipProb}; & \texttt{PrecipProb} > 0.72 \end{cases}$$

Model1 differs Model0 with the addition of `ppp`, which serves as a calibration term to `PrecipProb` to make it piecewise linear. The fitted parameter estimates, goodness-of-fit metrics are again summarized in Table 4.2. Each of the coefficients in Model1 is statistically significant at the $p < 0.01$ level. The adjusted R^2 also increases to 0.872 from 0.866 in Model0. The AIC and BIC both suggests a better fitting after accounting for the nonlinearity in `PrecipProb`. Counterfactual simulation is performed where the counterfactual X_c is created by varying `PrecipProb` while holding all other variables constant (i.e., set to their sample means). Figure 4.3 shows a clear inverse relationship between precipitation probability and bicycle counts. The rate of decrease in

bicycles appears to begin somewhat flat, and then begins to drop steeply at higher probability of precipitation. This suggests that bicyclists tend not to take the `PrecipProb` seriously in their riding decision until it is really high. This could be due to the constant presence of precipitation in Seattle and the inaccuracy of the weather report.

Another weather variable we are interested is the `Weather` variable. `Weather` is a general description of the weather condition and it is represented as a categorical variable in the model, such as 'fog', 'cloudy','partly-cloudy', etc. The hypothesis behind this is that people may make riding decisions based on the easily perceived weather classification rather than the actual temperature or precipitation number. The alternative model specification is as follows:

Model2

$$\sqrt{\text{Count}_t} = \beta_0 + \beta_1 \text{TempMax}_t^2 + \beta_2 \text{TempMax}_t + \beta_3 \text{Holiday}_t + \beta_4 \text{PrecpProb}_t + \beta_5 \text{Weekend}_t + \beta_6 \text{Daylight}_t + \beta_7 \text{Weather}_t + \beta_8 \text{UW}_t + \varepsilon. \quad (4.4)$$

Model2 differs Model0 with the addition of `Weather` variable. Model2's parameter estimates, goodness-of-fit metrics are summarized in Table 4.2. Noted that the weather condition 'fog', 'wind' and 'partly-cloudy' are found to be not statistically significant at $p < 0.1$ level. The increase in adjusted R^2 is also marginal given the addition of `Weather` variable. This suggests that adding `Weather` classifications variable doesn't have a significant impact on bike counts.

4.3 Seasonal Impact

In Model0, four variables are considered to the effect of seasonality on bicyclist counts: the first being the number of daylight hours `daylight`, the second being whether or not the University of Washington (proxying more generally for other educational institutions) was in session `UW`, third being the holiday indicator `Holiday`, and the fourth being the weekend indicator `Weekend`.

Figure 4.4 shows that we see a substantial increase in bicycle volume when the UW is in session. In this case, the counterfactual X_c is created by varying `UW` and `daylight` while holding all other variables constant (i.e., set to their sample means). With all other factors held the same we see that, on days when the university is in session, we would expect an average of approximately 367

Table 4.2: Results: Weather variables

	Dependent variable: $\sqrt{Bike\ Counts}$		
	Model0	Model1	Model2
TempMaxSq	-0.006^{***} (0.001)	-0.006^{***} (0.001)	-0.006^{***} (0.001)
TempMax	1.151^{***} (0.136)	1.179^{***} (0.133)	1.127^{***} (0.137)
Holiday	-14.680^{***} (0.850)	-14.277^{***} (0.829)	-14.809^{***} (0.840)
PrecipProb	-11.734^{***} (0.473)	-9.142^{***} (0.565)	-8.517^{***} (0.763)
ppp		-45.184^{***} (5.693)	
Weekend	-17.435^{***} (0.316)	-17.277^{***} (0.308)	-17.421^{***} (0.312)
daylight	1.016^{***} (0.088)	0.946^{***} (0.086)	1.037^{***} (0.090)
cloudy			-2.956^{***} (1.036)
fog			0.146 (0.702)
partly-cloudy-day			-0.107 (0.436)
partly-cloudy-night			0.579 (0.594)
rain			-2.785^{***} (0.561)
wind			-4.159 (3.404)
UW	3.272^{***} (0.348)	3.487^{***} (0.340)	3.266^{***} (0.344)
const.	-3.627 (3.796)	-4.775 (3.701)	-2.900 (3.818)
Observations	1,157	1,157	1,157
R^2	0.866	0.873	0.870
Adjusted R^2	0.866	0.872	0.869
AIC	6930	6870	6906
BIC	6976	6921	6982

Note: *p<0.1; **p<0.05; ***p<0.01

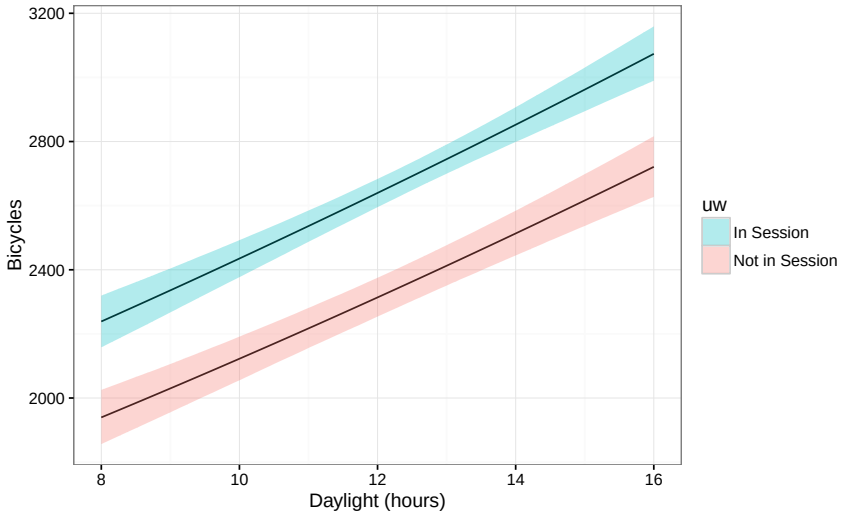


Figure 4.4: Effect of `Daylight` and `UW` in-session status on bicycle counts, all other factors held constant, with shaded 95% confidence interval.

additional bicycle observed. Similarly, we see a rougly linear increase in bicycles with increased day length.

Figure 4.5 shows the impact of weekend on the bike counts. In this case, only `Weekend` is varied while all other variables are held constant (i.e., set to their sample means). When the date of interest is a weekend, there is lower number of bicycle counts compared to a weekday. This strongly suggests that the majority of the bicycle traffic at this location is for utilitarian purpose.

Several additional alternative seasonal variables are considered for model specifications. In Model0, the continuously-valued variable `daylight` is considered to account for the yearly seasonality. In the following, the `daylight` variable is replaced by the traditional calendar season variable `season` taking values in 'spring', 'summer', 'fall' and 'winter'. The alternative model specification is given by:

Model3

$$\sqrt{\text{Count}_t} = \beta_0 + \beta_1 \text{TempMax}_t^2 + \beta_2 \text{TempMax}_t + \beta_3 \text{Holiday}_t + \beta_4 \text{PrecpProb}_t + \beta_5 \text{Weekend}_t + \beta_6 \text{Season}_t + \beta_6 \text{UW}_t + \varepsilon. \quad (4.5)$$

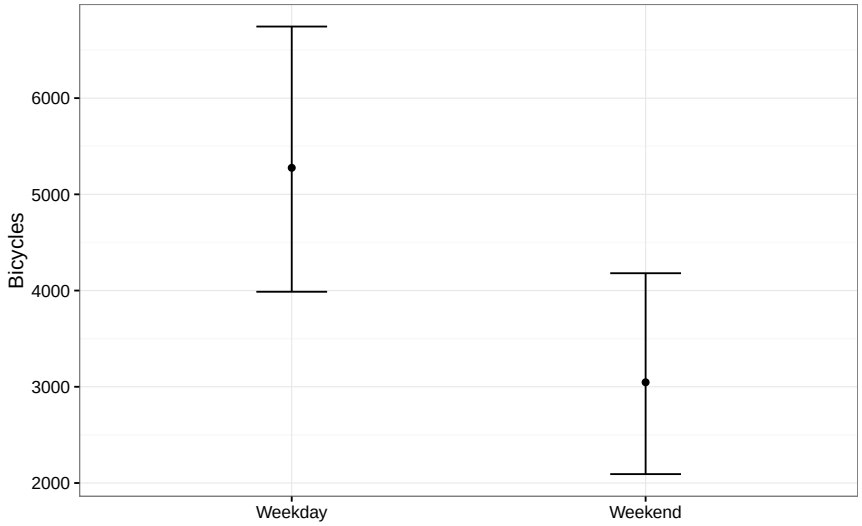


Figure 4.5: Effect of `weekend` on bicycle counts, all other factors held constant, with 95% confidence interval.

Model3 replaces the `Daylight` hours in Model0 by the `Season` variable. Model3's parameter estimates, goodness-of-fit metrics are summarized in Table 4.3. All coefficients of the season variables appear to be statistically significant in Model3 at the $p < 0.01$ level. The adjusted R^2 is slightly lower than Model0 suggesting using `daylight` gives a slightly better fit. Figure 4.6 shows that there is generally more bike counts in summer time than winter when holding all other variables constant. Considering that `dayligh` is a continuous-valued, it also avoids predicting big jumps on days such as March 31 and June 30 where calendar season changes. Therefore, the `daylight` is preferred in the master model.

Figure 4.7 capture the effect of `holiday`, `Weekend` and `Season` in one plot. The counterfactual is constructed by letting `holiday`, `Weekend` and `Season` vary while holding all other variables in Model3 (4.5) constant (i.e., set to sample means). It again shows in general, there is more bike volumes on weekday than weekends, summer than winter, and holiday than non-holiday. One interesting finding is that for weekdays that are also holidays, it is expected to have similar bicycle traffic as normal weekends.

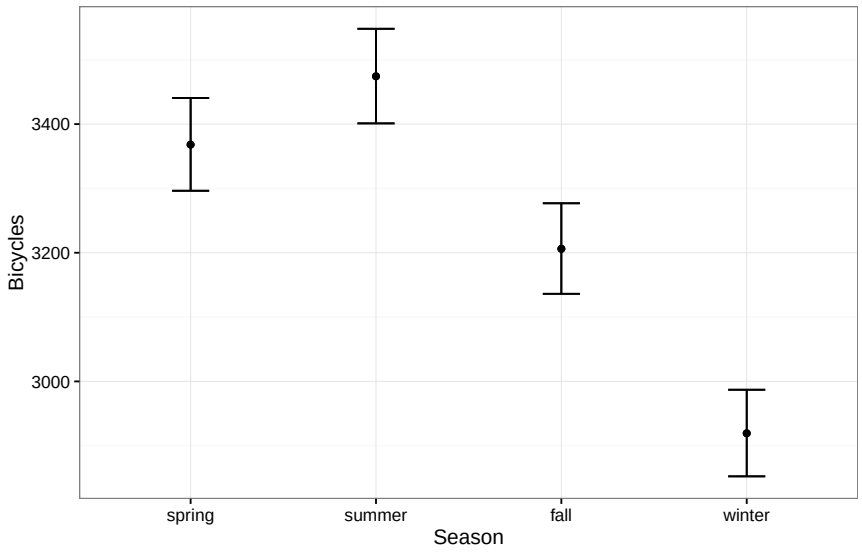


Figure 4.6: Effect of `Season` on bicycle counts, all other factors held constant, with 95% confidence interval.

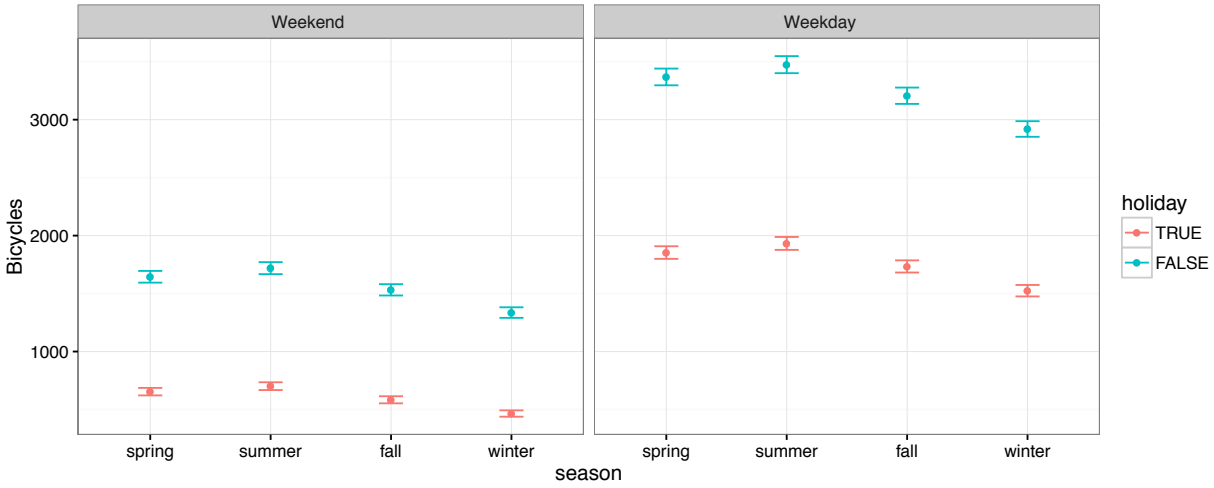


Figure 4.7: Effect of `holiday` on bicycle counts, all other factors held constant, with 95% confidence interval.

Next we consider the `Weekend` variable substituted by the day of week variable `dow`. The use of `dow` has been documented a lot in the literature [41, 4]. It is included as a categorical variable, such as 'Saturday', 'Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday' and 'Friday'. It is included to account for the weekly variation in bicycle counts apparent in timeseries plot. The alternative model specification is given by:

Model4

$$\sqrt{\text{Count}_t} = \beta_0 + \beta_1 \text{TempMax}_t^2 + \beta_2 \text{TempMax}_t + \beta_3 \text{Holiday}_t + \beta_4 \text{PrecpProb}_t + \beta_5 \text{dow}_t + \beta_6 \text{Daylight}_t + \beta_7 \text{UW}_t + \varepsilon. \quad (4.6)$$

Model4 replaces `Weekend` variable in Model0 by `dow` variable. Model4's parameter estimates, goodness-of-fit metrics are summarized in Table 4.3. Figure 4.8 shows several interesting aspects of these results. First, when holding all other variables in Model4 (4.6) constant (i.e., set to sample means), we see much higher number of bicyclists on weekdays than on weekends. Comparing the bike counts within different weekdays, we see that most weekdays have roughly the same bicycle traffic, with Tuesday having the most bike traffic and Friday the least. However, the information carried with `dow` is small (little difference between weekdays). Therefore, the `Weekend` is preferred in our model.

4.4 Interacting Relationship

Interaction effects among weather variables (e.g., combined effect of temperature and humidity) on the bike count have been studied in the literature [41]. However, few study is done on the interactions between weather variables and seasonal factors, although such interaction has been noticed in the literature [60]. In [60], it is concluded cycling on utilitarian facilities is more sensitive to weather conditions on weekends. In this alternative model specification, the interaction term between `Weekend` and `PrecipProb` is explicitly modelled:

Table 4.3: Results: Seasonal factors

	Dependent variable: $\sqrt{\textit{Bike Count}}$		
	Model0	Model3	Model4
TempMaxSq	-0.006^{***} (0.001)	-0.006^{***} (0.001)	-0.005^{***} (0.001)
TempMax	1.151^{***} (0.136)	1.167^{***} (0.132)	1.190^{***} (0.149)
Holiday	-14.680^{***} (0.850)	-14.424^{***} (0.827)	-14.985^{***} (0.874)
PrecipProb	-11.734^{***} (0.473)	-11.725^{***} (0.457)	-12.205^{***} (0.481)
Weekend	-17.435^{***} (0.316)		-17.483^{***} (0.324)
Sat		-14.271^{***} (0.514)	
Sun		-15.035^{***} (0.514)	
Mon		3.135^{***} (0.514)	
Tue		4.140^{***} (0.513)	
Wed		3.970^{***} (0.512)	
Thu		2.589^{***} (0.511)	
daylight	1.016^{***} (0.088)	1.016^{***} (0.085)	
winter			-2.592^{***} (0.492)
spring			1.413^{***} (0.417)
summer			2.320^{***} (0.543)
UW	3.272^{***} (0.348)	3.329^{***} (0.336)	3.242^{***} (0.384)
const.	-3.627 (3.796)	-6.969^{*} (3.692)	5.635 (4.491)
Observations	1,157	1,157	1,157
R^2	0.866	0.876	0.860
Adjusted R^2	0.866	0.875	0.859
AIC	6930	6855	6989
BIC	6976	6926	7045
Note:			*p<0.1; **p<0.05; ***p<0.01

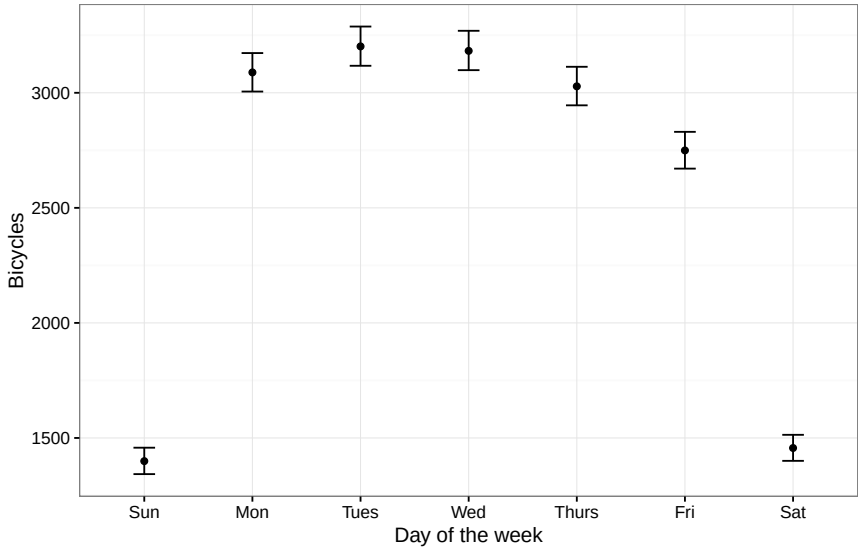


Figure 4.8: Effect of `dow` on bicycle counts, all other factors held constant, with 95% confidence interval.

Model5

$$\sqrt{\text{Count}_t} = \beta_0 + \beta_1 \text{TempMax}_t^2 + \beta_2 \text{TempMax}_t + \beta_3 \text{Holiday}_t + \beta_4 \text{PrecpProb}_t + \beta_5 \text{Weekend}_t + \beta_6 \text{PrecpProb}_t \times \text{Weekend}_t + \beta_7 \text{Daylight}_t + \beta_8 \text{UW}_t + \varepsilon. \quad (4.7)$$

Model5 differs from Model0 by allowing interactions between `PrecipProb` and `Weekend`. Model5's parameter estimates, goodness-of-fit metrics are summarized in Table 4.4. The coefficient for the interaction term is significant and with a negative value. This indicates that during weekends, the `PrecipProb` has a bigger negative effect on bike counts. This could be due to more bicyclists on weekends are recreational and thus more sensitive to weather conditions.

4.5 Trend Analysis

To test for a linear trend in bicycle volumes, we include a time index variable `day` in the model. We created this variable by sequentially numbering (1-1157) the observed counts by day during the study period. The alternative model specification is as follows:

Table 4.4: Results: Interaction and trend

	Dependent variable: $\sqrt{\textit{Bike Count}}$		
	Model0	Model5	Model6
TempMaxSq	-0.006^{***} (0.001)	-0.006^{***} (0.001)	-0.005^{***} (0.001)
TempMax	1.151^{***} (0.136)	1.163^{***} (0.136)	1.087^{***} (0.136)
Holiday	-14.680^{***} (0.850)	-14.663^{***} (0.848)	-14.690^{***} (0.845)
PrecipProb	-11.734^{***} (0.473)	-11.027^{***} (0.541)	-11.659^{***} (0.471)
Weekend	-17.435^{***} (0.316)	-16.722^{***} (0.413)	-17.441^{***} (0.314)
daylight	1.016^{***} (0.088)	1.015^{***} (0.088)	1.079^{***} (0.089)
UW	3.272^{***} (0.348)	3.263^{***} (0.347)	3.266^{***} (0.346)
PrecipProb:Wknd		-2.482^{***} (0.929)	
day			0.002^{***} (0.0004)
const.	-3.627 (3.796)	-4.144 (3.791)	-2.817 (3.780)
Observations	1,157	1,157	1,157
R^2	0.866	0.867	0.868
Adjusted R^2	0.866	0.866	0.867
AIC	6930	6925	6918
BIC	6976	6975	6968

Note: *p<0.1; **p<0.05; ***p<0.01

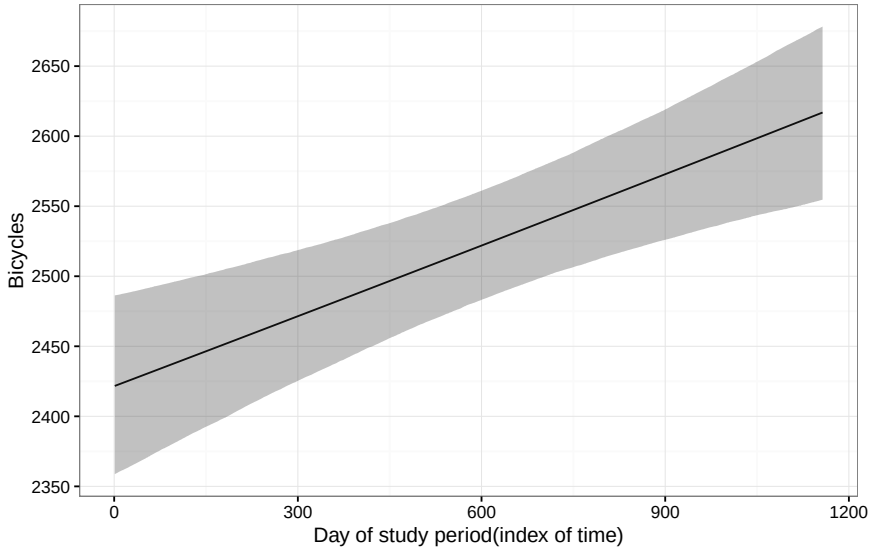


Figure 4.9: General trend in bicycling counts, all other factors held constant, with shaded 95% confidence interval.

Model6

$$\sqrt{\text{Count}_t} = \beta_0 + \beta_1 \text{TempMax}_t^2 + \beta_2 \text{TempMax}_t + \beta_3 \text{Holiday}_t + \beta_4 \text{PrecpProb}_t + \beta_5 \text{Weekend}_t + \beta_6 \text{daylight}_t + \beta_7 \text{UW}_t + \beta_8 \text{day}_t + \varepsilon. \quad (4.8)$$

Model6 differs from Model0 by the addition of `day`. Model6's parameter estimates, goodness-of-fit metrics are summarized in Table 4.4. Results shown in Figure 4.9 confirm the presence of a general trend toward increased numbers of bicycles at this location. With all else held constant, we would expect to see roughly 200 more bicycles (or 8%) of daily bike counts at the end of our 3-year study period than at the beginning. This is consistent with findings reported elsewhere that suggest that bicycling is increasing in a number of cities including Seattle [35]. However, if the growth rate stays the same for the next two decades, the goal of quadrupling cyclists by 2030 as proposed in Seattle Bicycle Master Plan [3] is going to fail. Therefore, this study suggests that the city of Seattle should take more aggressive actions to improve cycling experience and encourage cycling usage.

4.6 Master Model

Finally, we investigate the model specification which combines the results of all above exploration:

Model7 (Master Model)

$$\sqrt{\text{Count}_t} = \beta_0 + \beta_1 \text{TempMax}_t^2 + \beta_2 \text{TempMax}_t + \beta_3 \text{Holiday}_t + \beta_4 \text{PrecipProb}_t + \beta_5 \text{ppp}_t + \beta_6 \text{dow}_t + \beta_7 \text{Daylight}_t + \beta_8 \text{UW}_t + \beta_9 \text{Weather}_t + \beta_{10} \text{day} + \varepsilon. \quad (4.9)$$

When compared to Model0, Model7 includes the nonlinear terms `TempMax`^{2} and `ppp`, `dow`, `Weather` and the time index `day` to best capture the relationship between bike volumes and weather and seasonal factors. Model7's parameter estimates, goodness-of-fit metrics are summarized in Table 4.5. As expected, Model7 has the highest R-squared (0.887) and lowest AIC/BIC value (6755/6861) among all models. It suggests Model7 is able to explain about 89% of variation in the original observations. However, this comes with the cost of 10 explanatory variables. Considering model simplicity and the fact that the performance of Model0 is fairly close, Model0 will be chosen as our preferred model and will be used in prediction in Chapter 5.

Table 4.5: Results: Master model

	Dependent variable: $\sqrt{\textit{Bike Count}}$	
	Model0	Model7
TempMaxSq	-0.006^{***} (0.001)	-0.005^{***} (0.001)
TempMax	1.151^{***} (0.136)	1.115^{***} (0.129)
Holiday	-14.680^{***} (0.850)	-14.200^{***} (0.793)
PrecipProb	-11.734^{***} (0.473)	-6.292^{***} (0.762)
ppp		-42.413^{***} (5.416)
Weekend	-17.435^{***} (0.316)	
Mon		18.027^{***} (0.496)
Tues		19.053^{***} (0.491)
Wed		18.777^{***} (0.490)
Thur		17.429^{***} (0.491)
Fri		14.895^{***} (0.492)
Sat		0.752 (0.492)
daylight	1.016^{***} (0.088)	1.014^{***} (0.084)
UW	3.272^{***} (0.348)	3.522^{***} (0.323)
day		0.002^{***} (0.0004)
cloudy		-2.504^{**} (0.987)
fog		0.365 (0.658)
partly-cloudy		0.125 (0.363)
rain		-2.508^{***} (0.527)
wind		-3.550 (3.194)
intercept	-3.627 (3.796)	-21.593^{***} (3.572)
Observations	1,157	1,157
R^2	0.866	0.887
Adjusted R^2	0.866	0.886
AIC	6930	6755
BIC	6976	6861
Note:		*p<0.1; **p<0.05; ***p<0.01

Chapter 5

PREDICTIVE MODEL

In this chapter, we are interested in developing a predictive model for daily bike count using weather conditions and seasonal factors. Such model is useful to help transportation administrator to make better-informed decisions/preparations in case of inclement weather or special holidays. First, regression-based models are considered. Predictive performance of the four OLS regression-based models are compared. Root mean squared error is calculated and actual vs. predicted plot is also used to visualize the predication result. Next, a time series analysis approach is adopted to improve the prediction performance. An Autoregressive Integrated Moving Average (ARIMA) model is proposed to account for autocorrelation patterns that are observed in the data. The results of ARIMA is interpreted and its predictive performance is discussed.

5.1 Regression-based Model

We first use regression based models to predict daily bike volumes. The standard linear regression, log-linear model, square root transformation model and the Poisson model as described in 3.3.1 are tested. The Model0 specification (4.2) is used. To compare their predictive performance, an out-of-sample validation framework is used: each model is trained using the first two years' data, then the trained model is tested using the third year's bike count data. The root mean squared error (RMSE) is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(\hat{Y}_i - Y_i)^2}$$

where Y_i is the actual bike count at date i in the test data set, $\hat{Y}_i$ is the corresponding predicted bike counts, and N is the number of data points in the test data set. The root mean squared error of the four proposed models are reported in Table 5.1.

Name	RMSE
Standard linear regression	595.05
log-linear model ($\log Y$)	631.11
Square root transformation model ($\sqrt{Y}$)	575.89
Poisson model	585.32

Table 5.1: Predication performance on test data set in terms of RMSE, using the first two years' data as training data set

Test Name	$p-$value	Null Hypothesis
ADF test	0.01	Non-Stationary

Table 5.2: Statistical stationarity test with its $p-$value and null hypothesis

Out of the four models, the square root transformation model has the smallest RMSE of 575.89. Furthermore, the predicted bike volumes are relatively close to the actual data and most of the variation is predicted, as is shown in Figure 5.1, although there are some points that are under-predicted (the bottom right corner). While this error is perhaps too large for many operational uses, observed daily bike volume spans from 0 to 6000 (as is seen in Figure 5.1), making a prediction with this level of accuracy is quite useful for strategic decisions about network expansion.

5.2 Time Series Analysis with ARIMA

In this section, we seek to improve the predictive performance by adopting a time series analysis approach. More specifically, an ARIMA type model is proposed to account for the autocorrelation patterns observed. Following the Box-Jenkins approach [7], we first check the stationarity of the model residual errors. As is shown in the top left plot in Figure 5.2, there is no discernible long-run trend or seasonality in the residual responses. Furthermore, statistical stationarity test (i.e., Augmented Dickey-Fuller (ADF) test) is performed. The result is summarized in Table 5.2. Using a significance level of 0.05, ADF test suggests the time series are stationary.

The autocorrlation plot (ACF) and partial autocorrelation plot (PACF) are then used to exam-

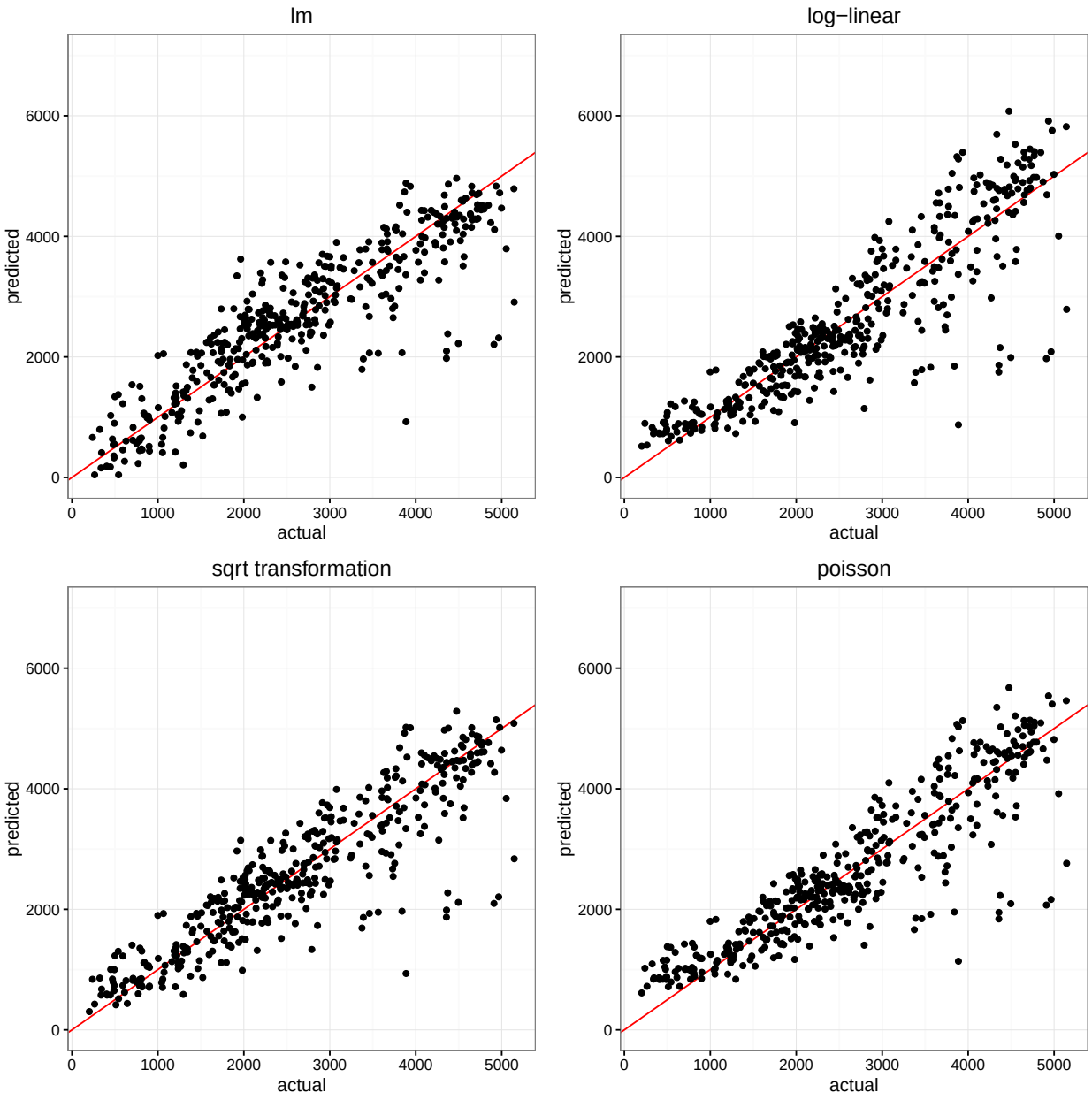


Figure 5.1: Actual vs. predicted daily bike volume for each of the four models in Table 5.1. The red line shows actual count is equal to predicted. This plot shows the majority variation is predicted, with some points under-predicted.

ine any possible autocorrelation, as are shown in Figure 5.2. Recall that an ACF plot shows the autocorrelation which measure the relationship between Y_t and Y_{t-k} for different values of k. The PACF measures the relationship between Y_t and Y_{t-k} after removing the effects of other time lags in between: $1, 2, 3, \ldots, k-1$. The ACF and PACF plots on the left side suggest there is significant autocorrelation in the residuals, indicated by the high spikes that do not decay.

Both the ACF and PACF plots of the residual responses (see middle and bottom left plots in Figure 5.2) have persistent spikes that do not disappear after certain lags, suggesting the order of autoregressive and moving average terms (p, q) are positive. In this case, the ACF and PACF are not helpful in finding the suitable values of p and q. To determine the appropriate order of ARIMA terms, we use the `auto.arima` function of `forecast` package in `R` to aid the decision. The result returned by `auto.arima` is $(2, 0, 1)$. Using $(2, 0, 1)$ as the base point, we implemented several variants of ARIMA models (by perturbing p and q in their respective neighbourhood) and compare their performance in terms of AIC value. It turns out that ARIMA(1,0,1) can achieve a very similiar result as ARIMA(2,0,1) with a simpler structure. Although complex ARIMA models are capable of producing models without correlation in model residual responses, it is our intention to keep simplicity of the predictive model so that: 1) it won't overfit; 2) it is easy to implement; 3) it has practical applications. For this reason, the ARIMA(1,0,1) model is chosen over ARIMA(2,0,1). The residual responses, ACF and PACF plots of ARIMA(1,0,1) is depicted in the right column of Figure 5.2. Incorporating ARIMA error structures clearly helps reduce the autocorrelation patterns in the residual errors compared to the original model.

Table 5.3 reports the estimated coefficients for both the ARIMA model and the base model (Square root transformation model without ARIMA error terms). The parameter estimates of the autoregressive term is statistically significant. The signs of other coefficients are generally the same in both models. However, the significance and magnitude of the coefficients are systematically larger in the base model (with the exception of `daylight`), indicating that the effects of weather variables are overestimated when complex serial correlation patterns in the error terms are not accounted for. This observation is consistent with the finding in [18]. The ARIMA model also has a larger R^2 and smaller AIC/BIC value, indicating a better fit of the data and superior predictive

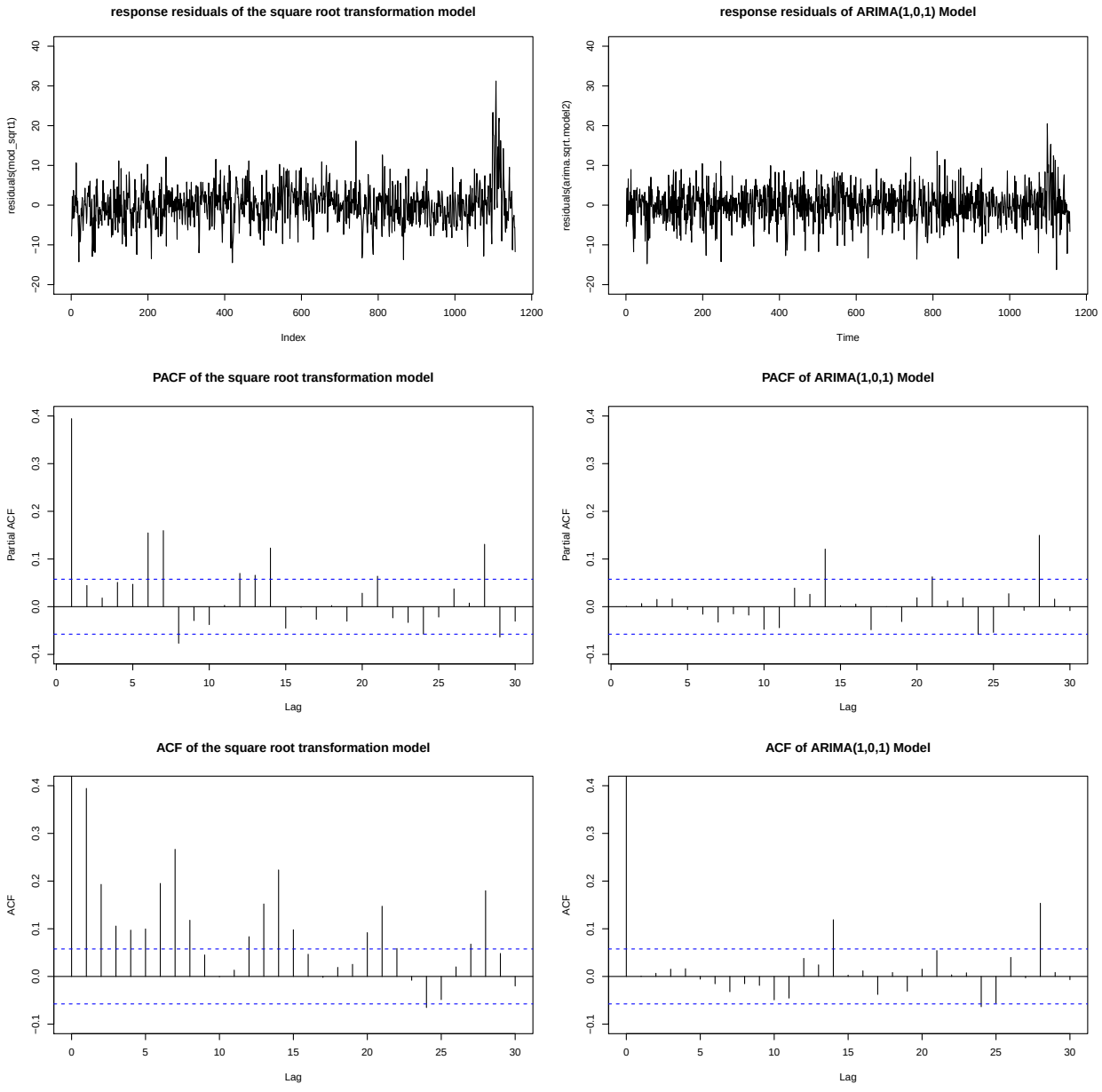


Figure 5.2: From top to bottom: the residual plots, ACF and PACF of the square root transformation model with ARIMA(1,0,1) error terms (right) and without ARIMA error terms (left).

performance.

Table 5.3: Square root transformation model with vs. without ARIMA error terms

	Dependent variable:sqrt(Bike Counts)	
	Square root transformation model	ARIMA(1,0,1)
TempMaxSq	-0.006^{***} (0.001)	-0.004^{***} (0.001)
TempMax	1.151^{***} (0.136)	0.845^{***} (0.172)
Holiday	-14.680^{***} (0.850)	-12.251^{***} (0.751)
PrecipProb	-11.734^{***} (0.473)	-10.459^{***} (0.475)
Weekend	-17.435^{***} (0.316)	-16.596^{***} (0.306)
UW	3.272^{***} (0.348)	2.666^{***} (0.532)
daylight	1.016^{***} (0.088)	1.256^{***} (0.131)
AR1		0.491^{***} (0.062)
MA1		-0.052 (0.072)
const.	-3.627 (3.796)	3.873 (4.929)
Observations	1,157	1,157
R^2	0.866	0.890
Adjusted R^2	0.866	0.889
Akaike Inf. Crit.	6986	6707
Bayesian Inf. Crit.	7045	6729
Note:		*p<0.1; **p<0.05; ***p<0.01

Last but not least, the predictive performance of the ARIMA model is evaluated through predicting the bike count for another year. The predicated daily bike counts for 2015, the third year in data series, are displayed in Figure 5.3. The actual (red) and forecasted (green) values are close

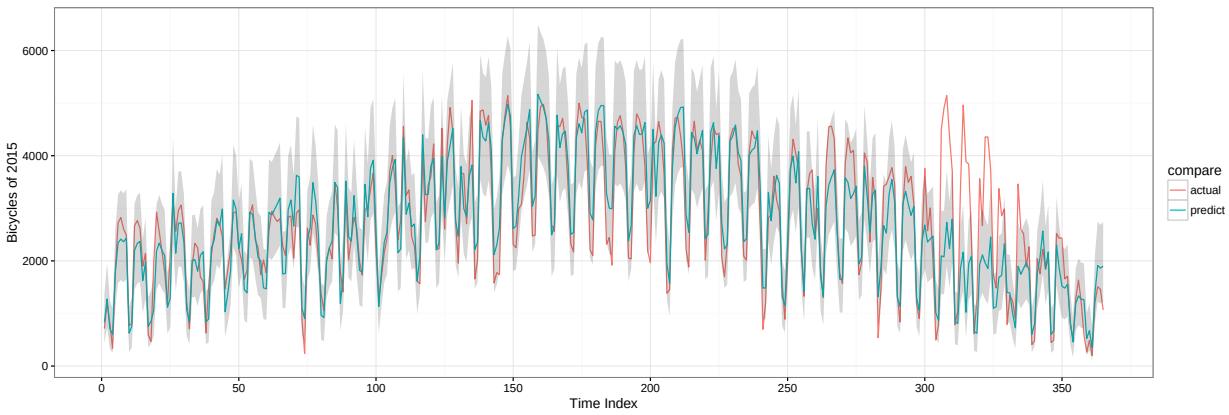


Figure 5.3: Plot of actual and predicted counts for 2015 test data with 95% predcition limits.

and the weekly and daily cycles generally match. The 95% prediction interval on the forecast are depicted as shaded area in Figure 5.3. For most of the time, the actual bicycle volume falls in the prediction interval, with the exception of a few days in November 2015. The actual bike counts has a surge which is even higher than the summer time. It could possibly due to a bike counter device malfunction. In general, the seasonal variation is clearly captured by the peak in summer and the bottom in winter.

Chapter 6

MULTI-LOCATION ANALYSIS

In this section, we extend our study in previous chapters to other bicycle locations with automatic counters in Seattle. The bike counters on different locations provide valuable information about the patterns of cycling in Seattle and allow more accurate benchmarking for goals to increase cycling as set forth in the Bicycle Master Plan. The goal of this section is to find out how different bike facilities respond to the same weather and seasonal variables, and identify their unique travel patterns. To achieve this, five bike facilities distributed in important bicycle network of Seattle are considered, and all of them have contiguous period of reliable data. To approximate a good representative of the bike behavior of Seattle, data in the past two years (from January 1, 2014 to December 31, 2015) are used in the model. Different patterns of the bike travel behavior are revealed by closely examining the data. Finally, a Poisson model is fitted for each location, and results are interpreted via simulation.

6.1 Locations Overview

This study utilizes cyclist count data from 5 automatic counting stations in the City of Seattle: Fremont Bridge, West Seattle Bridge, Elliott Bay Trail, I-90 Bridge and the Burke-Gilman Trail. These five locations are distributed in the existing Citywide Bicycle Network in different neighborhoods, which represents a network of all ages and abilities bicycle facilities with comfortable separation from motor vehicles. All five facilities are multi-use trail that allows two-way, off-street pedestrian and bicycle use [3]. They all equipped with the automatic, inductive loop bicycle counters to provide accurate bike counts. The locations are listed in Table 6.1 and shown in the map in Figure 6.1. Note that there seems to be three recreational facilities according to the official location description (Elliott, I-90 and BGT). However, by analyzing the data we will later find out the true

Table 6.1: Five Locations with bike counters

Name	Location
Fremont Bridge	Fremont Bridge
Seattle Bridge	South Spokane Street Bridge
Elliott	Elliott Bay Trail in Myrtle Edwards Park
I90	Mountains to Sound Trail west of I-90 Bridge
BGT	Burke Gilman Trail north of NE 70th St

travel pattern (whether recreational or utilitarian) for each of the five locations.

To shed light on the bicycling behavior, we first plot five sites' count data over the past two years. As is shown in Figure 6.2, the Fremont Bridge has a much higher daily bike traffic when compared to other facilities. This confirms the impression of Fremont Bridge being one of the busiest cycling locations in Seattle. The other four appear to have similar bike volumes. All five locations exhibit a seasonal pattern (peak in the summer and low in the winter) as well as day-to-day count variation. The daily variation is possibly due to different usage pattern of recreational and utilitarian cyclists. Figure 6.3 depicts the distribution of cycle counts for each of the five locations. Poisson model is used as a compromise between the five. It is also because it is simple and the ease to directly interpret.

6.2 Model Results

For each of the five locations, a Poisson model is fitted with same explanatory variables: `TempMax`, `TempMaxSq`, `Holiday`, `PrecipProb`, `Weekend`, `UW` and `Daylight`. Model results with coefficients estimates are summarized in Table 6.2. Recall that for a Poisson regression model $\log E[Y_i] = \alpha + X_i^T \beta$, one unit increase in X_i will result in $100(e^{\beta} - 1)\%$ increase in $E[Y_i]$.

Without exception, all the proposed explanatory variables have significant impact on the daily bike volumes. Precipitation has a clearly negative association with increased daily bike ridership. For I-90, with a 0.1 increase in `PrecipProb`, it is expected to have the largest 8.3% decrease in

Figure 6.1: Five bike counter facilities on map

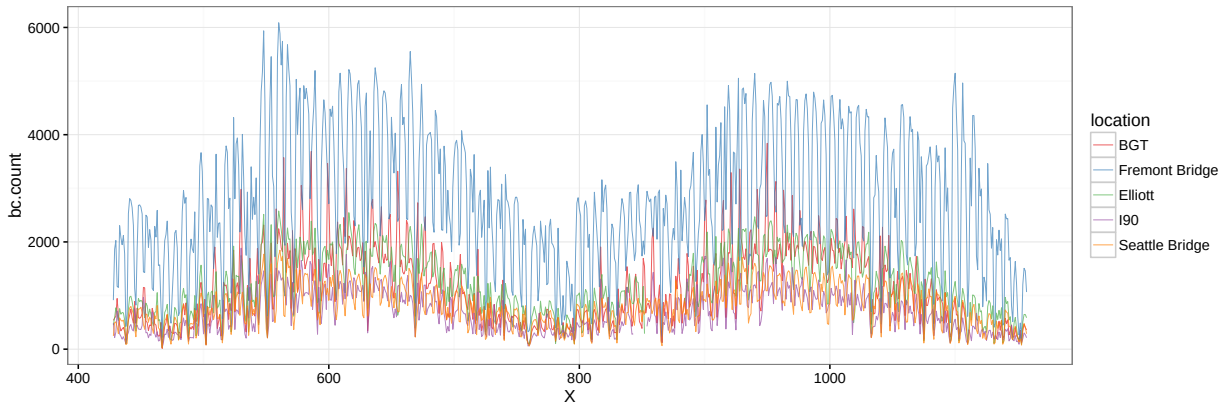


Figure 6.2: Bike counts for five locations

daily bike traffic, while for Fremont Bridge, the same amount of change only results in a 4% decrease. In terms of absolute number, BGT has the biggest drop for unit increase in the precipitation probability (as is shown in Figure 6.5-(a)), possibly due to its highly recreational usage.

The temperature variable has a nonlinear relationship with bike counts, which is indicated by the significance of coefficients for both `TempMax` and second order term `TempMaxSq`. With the help of counterfactual simulation, Figure 6.5-(b) depicts the nonlinear relationship between temperature and bike traffic. There is generally a positive impact on bike ridership when max temperature increases. However, when it is higher than 80 degree, there is a reduction in bike volumes for all five sites. Counterfactual simulations are further conducted to visualize the impact of max temperature on daily bike volume. The nonlinear 'leveling off' in counts at very high temperatures is well captured, as is shown in Figure 6.5-(b). In addition, the Burke-Gilman trail has the fastest declining rate possibly due to its recreational nature: recreational bikers tends to be more sensitive to high temperature than commuters.

The seasonal factor `Weekend` and `Holiday` provide a lot of insight to the nature of different bike facilities. It is shown that both `Weekend` and `Holiday` have positive impact on the daily bike traffic on Fremont Bridge, Seattle Bridge and Elliott Bay, and negative impact on I-90 and the Burke-Gilman Trail. This suggest that the former three are mainly utilitarian, while the latter two

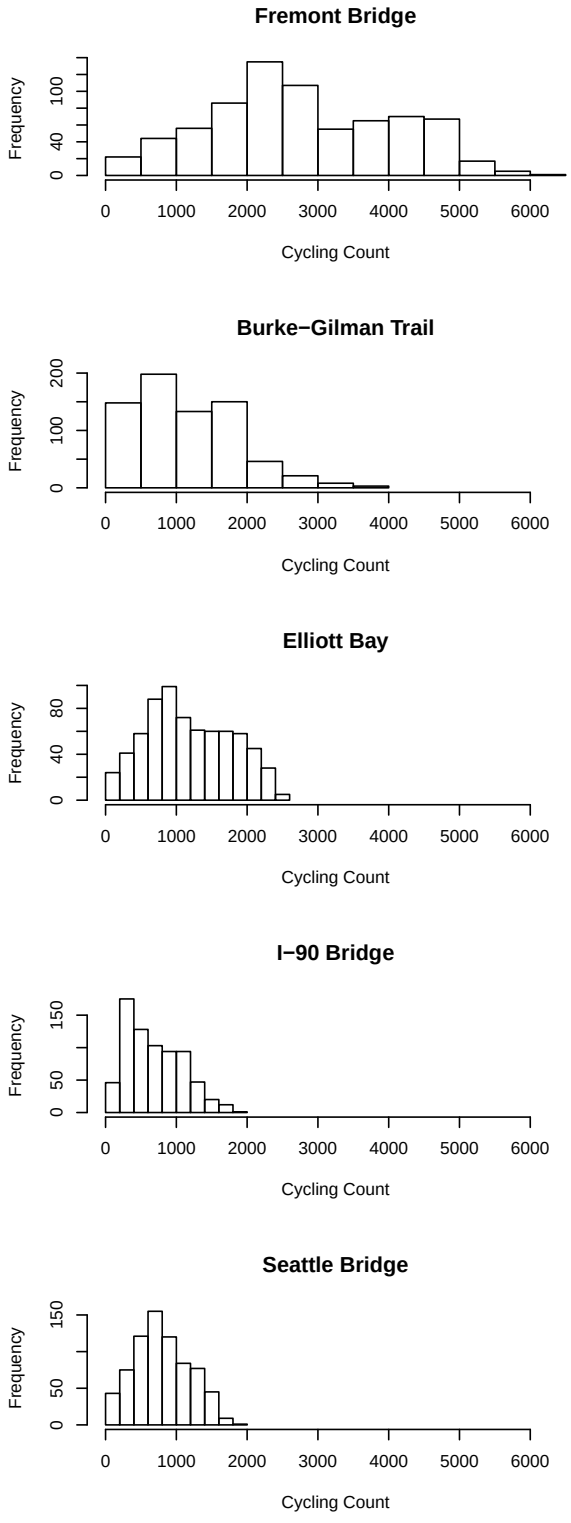


Figure 6.3: Distribution of bike counts at 5 locations

have more recreational users. However, according to the description in Table 6.1, Elliott Bay Trail was originally designed as a recreational facility. This finding from data indicates as time changes, more and more commuting cyclists now take route of the Elliott Bay trail and therefore change its traffic pattern. The Fremont Bridge is most sensitive to these two factors: there is 50% less bike traffic on weekends than weekdays, and 40% less on holidays than non-holidays. The absolute number of difference in bike counts is approximately 1600 and 1200 respectively, as suggested by Figure 6.6 and Figure 6.7. This is consistent with the fact that Fremont Bridge serves as a major corridor connecting northern Seattle with downtown Seattle. Another interesting findings is that Fremont not only have the highest amount of commuting bicyclists during weekdays, it also has arguably the largest recreational traffic during weekends and holidays. Also, since the Burke-Gilman trail is directly connected, they have similiar average amount of bike counts at weekends and holidays when other factors are held the same. In effect, the Fremont Bridge becomes a recreational facility during weekends and holidays.

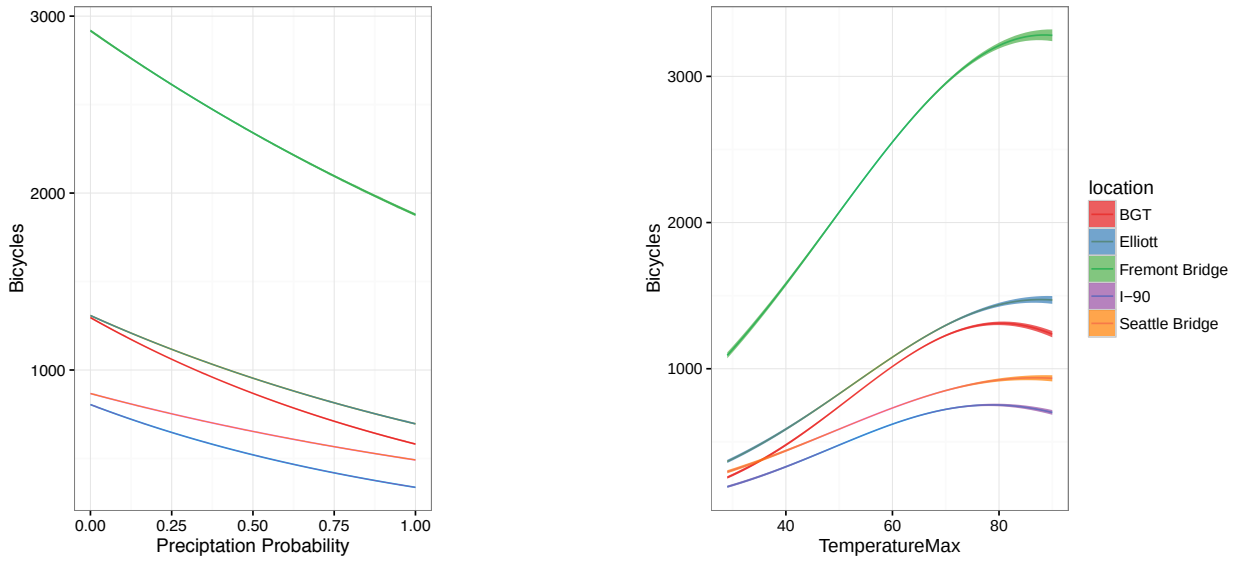
Other seasonal factors such as UW in session (`UW`) and daylight hour (`Daylight`) both have a positive impact on bike traffic for all five sites. As is shown in Figure 6.8, there is not much difference in absolute bike counts for BGT, Elliott Bay, I-90 and Seattle Bridge whether UW is in session or not. However, there is a significant increase in bike counts on the Fremont Bridge when UW is in session. This suggests that a lot of UW students/staff use Fremont Bridge to commute from and to school. At the same time, BGT is most sensitive to the daylight length, showing a 8% increase in bike counts with 1 hour increase in daily daily light hour. In addition, we also try to use the categorical substitute `Season` for `daylight`, which represents the more traditional calendar-based seasonality notion. It is observed in Figure 6.10 that bike counts on Fremont bridge peaks in summer and decreases significantly in winter. Also, there are more people riding bikes during spring and summer than fall and winter across five locations.

Goodness-of-fit results are also provided in Table 6.2 in terms of AIC, log-likelihood and deviance. Since the five sites have different datasets and traffic levels, model performance across different sites cannot be directly compared.

Table 6.2: Location Comparison

	Fremont Bridge	Seattle Bridge	Elliott	I-90	BGT
const.	5.49***	3.83***	3.70***	2.30***	2.09***
	(0.02)	(0.04)	(0.04)	(0.05)	(0.04)
TempMaxSq	−0.00***	−0.00***	−0.00***	−0.00***	−0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
TempMax	0.06***	0.06***	0.07***	0.09***	0.10***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Holiday	−0.55***	−0.46***	−0.31***	0.14***	0.32***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PrecipProb	−0.44***	−0.57***	−0.63***	−0.87***	−0.80***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Weekend	−0.75***	−0.62***	−0.43***	0.30***	0.36***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
UW	0.04***	0.05***	0.06***	0.08***	0.09***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Daylight	0.12***	0.12***	0.11***	0.12***	0.08***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
AIC	85717.93	29145.00	42591.48	33898.75	50548.60
Log Likelihood	-42850.96	-14564.50	-21287.74	-16941.37	-25266.30
Deviance	78679.17	23026.06	36464.50	27994.75	44399.89
Num. obs.	730	730	699	720	707

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$



(a) (b)

Figure 6.5: Effect of precipitation (a) and temperature (b) on counts for five locations, with shaded 95% confidence interval.

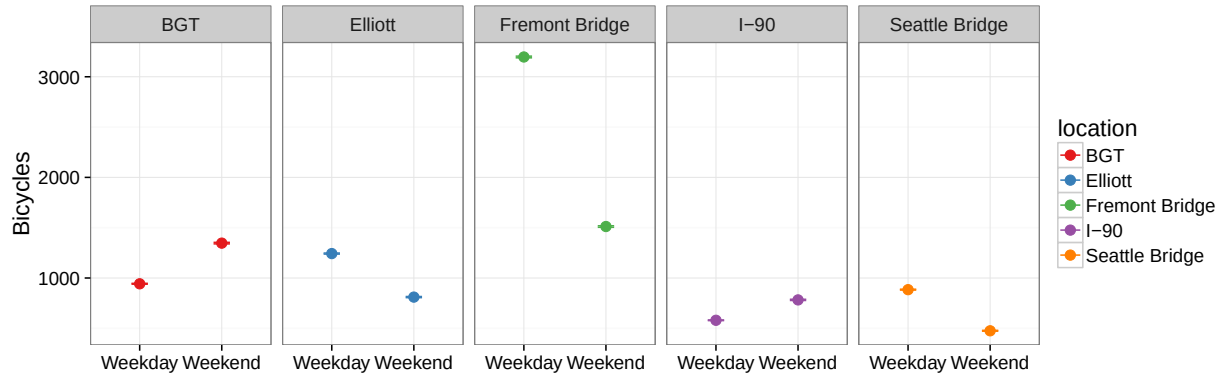


Figure 6.6: Effect of weekend on counts for five locations, with 95% confidence interval.

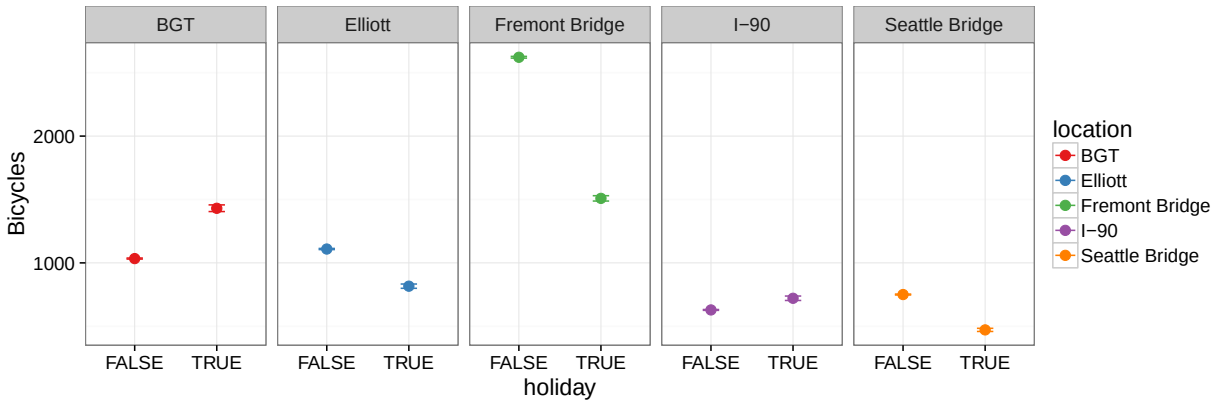


Figure 6.7: Effect of holiday on counts for five locations, with 95% confidence interval.

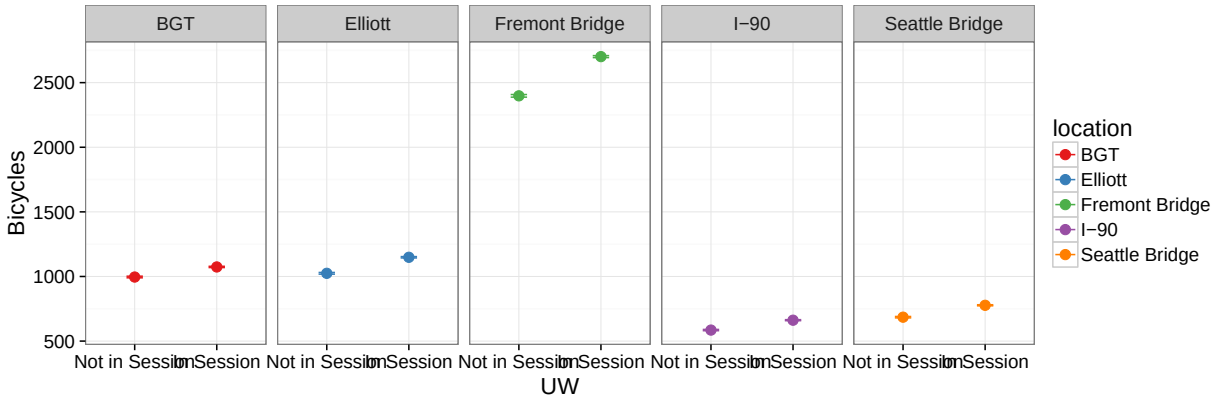


Figure 6.8: The impact of `UW` in session on counts for 5 locations, with 95% confidence interval.

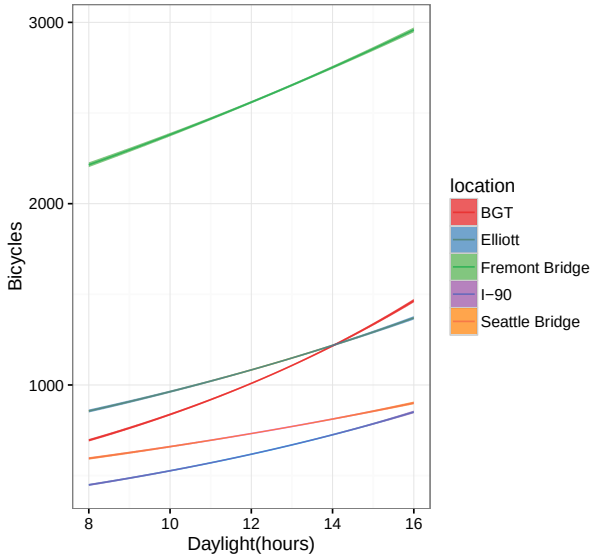


Figure 6.9: The Impact of daylight hour on counts for 5 locations, with shaded 95% confidence interval.

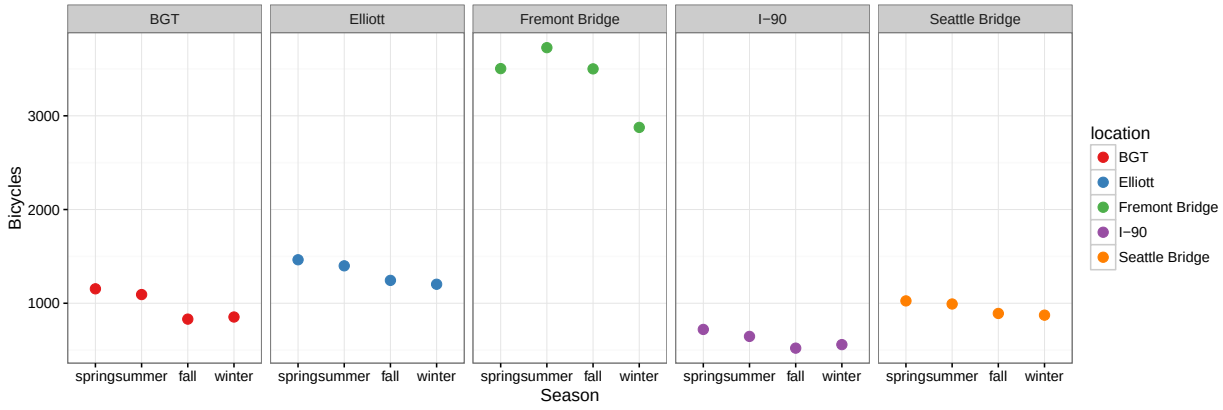


Figure 6.10: The Impact of calendar-based seasons on counts for 5 locations.

Chapter 7

CONCLUSION

Cycling is now becoming an increasingly important mode in a comprehensive transportation system for its physical, social, environmental and economic benefits. The city of Seattle has a vision to design and implement bicycle facilities that are safe and friendly for riders of all ages and abilities. To achieve this goal, it's crucial for the government stakeholders, policy makers and transportation planners to understand the factors that could drive the bike demand and influence bike travel behaviour. This thesis seeks to contribute to our limited knowledge on this topic.

The first contribution of this work is to thoroughly investigate the relationship between daily bicycle ridership with a set of key variables, including weather conditions and seasonal factors. The bike count data is continuously collected at the Fremont Bridge in Seattle for the past three years. It is joined with the daily weather conditions and other seasonal variables to form a rich modeling data. A systematic analysis is conducted to identify candidate explanatory variables and their appropriate forms of transformation that should be included in the final model. A regression model is then proposed to adequately capture the relationship between various factors and bike counts. This study, like similiar studies on this topic, confirms that weather conditions and seasonal factors have important impact on bike volumes. The maximum daily temperature and precipitation probability have the most significant effect on bicycle traffic counts. Holiday and weekend indicators are found to have negative impact on daily ridership. This could be explained by the fact that the bike facility under consideration is mainly utilitarian and less bike traffic is expected during holidays and weekend. With UW in session, more bike traffic will be expected possibly due to the facility's proximity to the University of Washington. Daylight hour, which serves as an indicator of the seasonality, takes account for the variations observed in different seasons. The model coefficients are reported in Table 4.2-4.4, as well as visually interpreted in Figure 4.1-4.9.

One of the major contributions of this work however lies in the discussion on nonlinear relationship between temperature/precipitation and bike count. It is explicitly modelled and accounted for using the general additive mixture (GAM) model. The squared temperature, which is commonly used in the literature, can be straightforwardly justified by the smooth function of temperature obtained by the GAM model. By incorporating the high order non-linear terms, the leveling off in counts at high temperature is well captured as is shown in the counterfactual simulations Figure 4.2. In addition, a piecewise linear model for the precipitation probability is recommended by the GAM model. It suggests that utilitarian bikers commuting on Fremont Bridge tend to be insensitive to rainfall unless the raining probability is very high.

With the exception of [18, 45], no other known research has studied the autocorrelation patterns in the bike count data. This research offers an ARIMA-type time series analysis model to account for the significant autocorrelation identified in the daily bike volumes. Note that there are previous works that have considered the hourly autocorrelation case in [18, 45]. The proposed model is used to predict future bike volumes and its predictive performance is evaluated using the train-test framework. The proposed ARIMA model could be used to assist planners by predicting potential bike traffic fluctuation in case of inclement weather, therefore better accommodate the choices cyclists make under such conditions.

Last but not least, a multi-location study is conducted to investigate the seasonal patterns and the effect of weather across five bike facilities in Seattle - Fremont Bridge, Seattle Bridge, Elliott Bay, I-90 Bridge, and the Burke-Gilman trail, which consists of both utilitarian and recreational purpose facilities. Results were generally in accordance with prior research on the Fremont Bridge. Weather conditions, such as temperature and precipitation, have similiar effect on all five sites, although the recreational facilities are shown to be more sensitive to weather conditions than utilitarian ones. Seasonal factors, however, such as weekend and holiday indicators, show distinct impact patterns on utilitarian and recreational facilities. Such findings can be used to help determine the composition of users of a particular facilities and therefore provide better infrastructure planning and program design.

Future research directions include a further investigation on the relationship between built envi-

ronment and bike ridership. Factors such as facility accessibility and bike network connectivity are expected to have significant impact on people's choice of cycling. A study on this could potentially shed light to future planning of the bike facilities and funding allocation. Also, while control of the weather and seasons are admittedly beyond the scope of policy makers, this research does suggest that planners and policy makers may want to develop strategies that help mitigate the impacts of the natural environment during the winter months. In other words, the delta between warm dry days and cold wet days should be treated as the opportunity frontier. Future research could focus on determining what, if any, programmatic or built interventions could ameliorate unfavorable cold- and wet-weather bicycling conditions.

BIBLIOGRAPHY

[1] Commute seattle survey finds increased transit to downtown, 2013.

[2] Conducting bicycle and pedestrian counts: A manual for jurisdictions in los angeles county and beyond. Technical report, The Southern California Association of Governments, Los Angeles County Metropolitan Transportation Authority, June 2013.

[3] Seattle bicycle master plan: Implementation plan 2015-2019. Technical report, The Seattle Department of Transportation, March 2015.

[4] F. Ahmed, Rose, G., M. Figliozzi, and C. Jakob. Commuter cyclist?s sensitivity to changes in weather: Insight from two cities with different climatic conditions. *Transportation Research Board Annual Meeting*, (2233), 2012.

[5] D. Asteriou and S. G. Hall. Arima models and the boxjenkins methodology. *Applied Econometrics*, pages 265–286, 2011.

[6] Alon Bassok, Noa Ginger, Andy Hong, John Murphy, Danielle Rose, Peter Schmiedeskamp, Amanda Snypp, and Eiji Torikai. Bicycle planning, best practices, and count methodology. Technical report, Puget Sound Regional Council and University of Washington Department of Urban Design and Planning, Seattle, 2011.

[7] G. Box and G. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1970.

[8] G. E. P. Box and D. R. Cox. An analysis of transformations. *J. Roy. Statist. Soc. B*, 26:211–252, 1964.

[9] C. Brandenburg, A. Matzarakis, and Arnberger A. Weather and cycling a first approach to the effects of weather on cycling. *Meteorol. Appl.*, 14, 2007.

[10] City of Seattle. Fremont bridge bike counter.

[11] City of Seattle. Open data.

[12] G.A. Colditz, C.C. Cannuscio, and A. L. Frazier. Physical activity and reduced risk of colon cancer: implications for prevention. *Cancer Causes Control*, 8:649–667, 1997.

[13] Alex Couture-Beil. *rjson: JSON for R*, 2014. R package version 0.2.15.

[14] J. Dill and K. Voros. Factors affecting bicycling demand: Initial survey findings from the portland, oregon. *Transportation Research Record*, 2031:9–17, 2007.

[15] M. K. Dunlap. How the weather, land use, and infrastructure influence non-motorized mode choices. Master's thesis, University of Washington, 2015.

[16] D. J. Fagnant and K. Kockelman. A direct-demand model for bicycle counts: the impacts of level of service and other factors. *Environment and Planning B: Planning and Design*, 43(1):93–107, 2016.

[17] Billy Fields. Active transportation measurement and benchmarking development: New orleans state of active transportation report 2010. Paper 4, UNOTI Publications, 2012.

[18] Christopher Gallop, Cindy Tse, and Jinhua Zhao. A seasonal autoregressive model of vancouver bicycle traffic using weather variables. *Transportation Research Board 2012 Annual Meeting*, 2012.

[19] J. Garrard, G. Rose, and S. K. Lo. Promoting transportation cycling for women: The role of bicycle infrastructure. *Preventative Medicine*, 46(1):55–59, 2008.

[20] Julia B. Griswold, Aditya Medury, and Robert J. Schneider. Pilot models for estimating bicycle intersection volumes. Institute of transportation studies, research reports, working papers, proceedings, Institute of Transportation Studies, UC Berkeley, 2011.

[21] Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011.

[22] Gung. Interpreting plot.lm().

[23] Z. Guo, N. H. M. Wilson, and A. Rahbee. Impact of weather on transit ridership in chicago, illinois. *Transportation Research Record: Journal of the Transportation Research Board*, (2034):3–10, 2007.

[24] S. Hanson and P. Hanson. Evaluating the impact of weather on bicycle use. *Transportation Research Record: Journal of the Transportation Research Board*, (629):43–48, 1977.

[25] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.

[26] Marco Helbich, Lars Böcker, and Martin Dijst. Geographic heterogeneity in cycling under various weather conditions: evidence from greater rotterdam. *Journal of Transport Geography*, 38(0):38 – 47, 2014.

[27] J.D. Hunt and J.E. Abraham. Influences on bicycle use. *Transportation*, 34(4):453–470, 2007.

[28] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. Business & Economics, 2013.

[29] Michael Jones and Lauren Buckland. Estimating Bicycle and Pedestrian Demand in San Diego. Technical report, UC Berkeley Safe Transportation Research & Education Center, Caltrans Task Order 6117, Berkeley, 2008.

[30] Michael G. Jones, Sherry Ryan, Jennifer Donlon, Lauren Ledbetter, David R. Ragland, and Lindsay Arnold. Seamless Travel: Measuring Bicycle and Pedestrian Activity in San Diego County and its Relationship to Land Use, Transportation, Safety, and Facility Type. Technical report, UC Berkeley Safe to Research & Education Center, Caltrans Task Order 6117, Berkeley, June 2010.

[31] A. Khattak and A. Palma. The impact of adverse weather conditions on the propensity to change travel decisions: A survey of brussels commuters. *Transportation Research Part A*, 31(3):181–203, 1997.

[32] J. K. Kim, S. Kim, G. F. Ulfarsson, and L. A. Porrello. Bicyclist injury severities in bicyclemotor vehicle accidents. *Accident Analysis and Prevention*, 39:238–251, 2007.

[33] K. J. Krizek, G. Barnes, and K. Thompson. Analyzing the effect of bicycle facilities on commute mode share over time. *Journal of Urban Planning and Development*, 135(2):66–73, 2009.

[34] Duncan Temple Lang. *RCurl: General network (HTTP/FTP/...) client interface for R*, 2014. R package version 1.95-4.3.

[35] League of American Bicyclists. Updated: Bike commute data released.

[36] I. Lee and P. Skerett. Physical activity and all-cause mortality: what is the dose-response relation? *Med Sci Sports Exerc*, 33:459–471, 2001.

[37] Amy Lewin. Temporal and weather impacts on bicycle volumes. *Transportation Research Board 90th Annual Meeting*, 2011.

[38] T Litman. *Evaluating transportation equity: Guidance for incorporating distributional impacts in transportation planning*. Victoria Transport Policy Institute, 2007.

[39] Chris McCahil. The applicability of space syntax to bicycle facility planning. *Transportation Research Record: Journal of the Transportation Research Board*, 2074(1):46 – 51, 2008.

[40] A.A. McDonald and New Zealand. Land Transport NZ. *Estimating Demand for New Cycling Facilities in New Zealand*. Land Transport New Zealand Research Report Series. New Zealand Government - NZ Transport Agency, 2007.

[41] L. F. Miranda-Moreno and Thomas N. Weather or not to cycle; whether or not cyclist ridership has grown: A look at weather's impact on cycling facilities and temporal trends in an urban environment. *Transportation Research Board 2011 Annual Meeting*, 2011.

[42] M. Nankervis. The effect of weather and climate on bicycle commuting. *Transportation Research Part A*, 33:417–431, 1999.

[43] Debbie A Niemeier. Longitudinal analysis of bicycle count variability: Results and modeling implications. *Journal of Transportation Engineering*, 122(3):200–206, 1996.

[44] K. Nordback, W. E. Marshall, B. N. Janson, and E. Stolz. Estimating annual average daily bicyclists: Error and accuracy. *92nd Annual Meeting of the Transportation 489 Research Board*, 2013.

[45] Thomas Nosal and Luis F. Miranda-Moreno. The effect of weather on the use of north american bicycle facilities: A multi-city analysis using automatic counts. *Transportation Research Part A: Policy and Practice*, 66(0):213 – 225, 2014.

[46] John Parkin, Mark Wardman, and Matthew Page. Estimation of the determinants of bicycle mode share for the journey to work using census data. *Transportation*, 35(1):93–109, 2008.

[47] J. Phung and G. Rose. Temporal variations in usage of melbourne's bike paths. *proceedings of 30th Australasian Transport Research Forum*, pages 25–27, 2007.

[48] Abdul Rawoof Pinjari, Chandra R. Bhat, and David A. Hensher. Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B: Methodological*, 43(7):729 – 748, 2009.

[49] J. Pucher, J. Dill, and S. Handy. Infrastructure, programs, and policies to increase bicycling: an international review. *Preventive Medicine*, 50:106–125, 2010.

[50] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[51] A. J. Richardson. Seasonal and weather impacts on urban cycling trips. *TUTI Report*, 2000.

[52] A. J. Richardson. Estimating bicycle usage on a national cycle network. *Transportation Research Record*, (1982):166–173, 2006.

[53] G Rose, F Ahmed, M Figliozzi, and C Jakob. Quantifying and comparing the effects of weather on bicycle demand in melbourne (australia) and portland (usa). *TRB 90th Annual Meeting:*, 2011.

[54] M. Schade. Explore the links between weather and capital bikeshare ridership.

[55] P. Schmiedeskamp and W. Zhao. Estimating daily bicycle counts in seattle from seasonal and weather factors. *Transportation Research Record: Journal of the Transportation Research Board*, To Appear 2016.

[56] W. L. Schwartz, Cambridge Systematics., Bicycle Federation of America., and Turner-Fairbank Highway Research Center. *Guidebook on methods to estimate non-motorized travel [microform] : supporting documentation / [authors, W.L. Schwartz ... et al.]*. U.S. Dept. of Transportation, Federal Highway Administration, Research, Development, and Technology, Turner-Fairbank Highway Research Center ; Available to the public through the National Technical Information Service McLean, VA (6300 Georgetown Pike, McLean 22101-2296) : [Springfield, Va, 1999.

[57] I. N. Sener, N. Eluru, and C. R. Bhat. An analysis of bicycle route choice preferences in texas. *Transportation*, 36(5):511–539, 2009.

[58] The Dark Sky Company. Forecast api docs.

[59] T. Thomas, R. Jaarsma, and B. Tutert. Temporal variations of bicycle demand in the netherlands: Influence of weather on cycling. *CD Proceedings of the 88th Annual Meeting of the Transportation Research Board*, 2009.

[60] T. Thomas, R. Jaarsma, and B. Tutert. Exploring temporal fluctuations of daily cycling demand on dutch cycle paths: the influence of weather on cycling. *Transportation*, 40, 2012.

[61] Sandar Tin Tin, Alistair Woodward, Elizabeth Robinson, and Shanthi Ameratunga. Temporal, seasonal and weather effects on cycle volume: an ecological study. *Environmental Health*, 11(1):12, 2012.

[62] S. Turner, A. Roozenburg, and T. Francis. Predicting accident rates for cyclists and pedestrians. *Land Transport New Zealand*, 289, 2006.

[63] S. Turner, G. Shunk, A. Hottenstein, Texas. Dept. of Transportation. Office of Research, Technology Transfer, and Texas Transportation Institute. *Development of a methodology to estimate bicycle and pedestrian travel demand*. Research report (Texas Transportation Institute). Texas Transportation Institute, Texas A&M University System, 1998.

[64] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.

[65] M. Winters, M. C. Friesen, M. Koehoorn, and K. Teschke. Utilitarian bicycling: A multilevel analysis of climate and personal influences. *American Journal of Preventive Medicine*, 32, 2007.

[66] R. Wittink. Planning for walking and cycling in urban environments. *In Sustainable transport*, 8:172–188, 2003.

[67] Y. Xing, S. L. Handy, and P. Mokhtarian. Factors associated with proportions and miles of bicycling for transportation and recreation in six small u.s. cities. *Transportation Research Part D*, 14(2):73–81, 2010.

[68] Schwartz Z Yu, G and J Walsh. A weather-resolving index for assessing the impact of climate change on tourism related climate resources. *Climate Change*, 95, 2009.