STRATEGIC DESIGN OF SMART BIKE-SHARING SYSTEMS FOR SMART CITIES

Huthaifa Issam Ashqar

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
In
Civil Engineering

Hesham A. Rakha, Chair
Leanna L. House, Co-Chair
Montasir M. Abbas
Hao Yang

September 21st, 2018
Blacksburg, VA

Keywords: Bike-Sharing System, Quality-of-Service, Transportation Mode Recognition, Urban Computing, Big Data.

Strategic Design of Smart Bike-Sharing Systems for Smart Cities

Huthaifa Issam Ashqar

ABSTRACT

Traffic congestion has become one of the major challenging problems of modern life in many urban areas. This growing problem leads to negative environmental impacts, wasted fuel, lost productivity, and increased travel time. In big cities, trains and buses bring riders to transit stations near shopping and employment centers, but riders then need another transportation mode to reach their final destination, which is known as the *last mile* problem. A smart bike-sharing system (BSS) can help address this problem and encourage more people to ride public transportation, thus relieving traffic congestion.

At the strategic level, we start with proposing a novel two-layer hierarchical classifier that increases the accuracy of traditional transportation mode classification algorithms. In the transportation sector, researchers can use smartphones to track and obtain information of multi-mode trips. These data can be used to recognize the user's transportation mode, which can be then utilized in several different applications; such as planning new BSS instead of using costly surveys. Next, a new method is proposed to quantify the effect of several factors such as weather conditions on the prediction of bike counts at each station. The proposed approach is promising to quantify the effect of various features on BSSs in cases of large networks with big data. Third, these resulted significant features were used to develop state-of-the-art toolbox algorithms to operate BSSs efficiently at two levels: network and station. Finally, we proposed a quality-of-service (QoS) measurement, namely *Optimal Occupancy*, which considers the impact of inhomogeneity in a BSS. We used one of toolbox algorithms modeled earlier to estimate the proposed QoS. Results revealed that the Optimal Occupancy is beneficial and outperforms the traditionally-known QoS measurement.

Strategic Design of Smart Bike-Sharing Systems for Smart Cities

Huthaifa Issam Ashqar

GENERAL AUDIENCE ABSTRACT

A growing population, with more people living in cities, has led to increased pollution, noise, congestion, and greenhouse gas emissions. One possible approach to mitigating these problems is encouraging the use of bike-sharing systems (BSSs). BSSs are an integral part of urban mobility in many cities and are sustainable and environmentally friendly. As urban density increases, it is likely that more BSSs will appear due to their relatively low capital and operational costs, ease of installation, pedal assistance for people who are physically unable to pedal for long distances or on difficult terrain, and the ability to track bikes in some cases.

This dissertation is a building block for a smart BSS in the strategic level, which could be used in real and different applications. The main aims of the dissertation are to boost the redistribution operation, to gain new insights into and correlations between bike demand and other factors, and to support policy makers and operators in making good decisions regarding planning new or existing BSS.

This dissertation makes many significant contributions. These contributions include novel methods, measurements, and applications using machine learning and statistical learning techniques in order to design a smart BSS. We start with proposing a novel framework that increases the accuracy of traditional transportation mode classification algorithms. In the transportation sector, researchers can use smartphones to track and obtain information of multi-mode trips. These data can be used to recognize the user's transportation mode, which can be then used in planning new BSS. Next, a new method is proposed to quantify the effect of several factors such as weather conditions on the prediction of bike station counts. Third, we use state-of-the-art data analytics to develop a toolbox to operate BSSs efficiently at two levels: network and station. Finally, we propose a quality-of-service (QoS) measurement, which considers the impact of inhomogeneity of BSS properties.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Hesham Rakha for his guidance and assistance, which led me to the right direction of this research and encouraged me to reach this great milestone in my life. I would like also to express my gratitude to my co-advisor, Dr. Leanna House for her guidance and effort. I am very thankful to my committee members; Dr. Montasir Abbas and Dr. Hao Yang for their support and illuminating feedback. I really do appreciate their time and invaluable academic advices which enhanced my knowledge and research.

I dedicate this work to my parents, Prof. Issam and Wafaa; my sister, Walaa; my brothers; and my wife, Nour.

TABLE OF CONTENTS

ABSTRACT ii
GENERAL AUDIENCE ABSTRACT iii
ACKNOWLEDGEMENTS iv
TABLE OF CONTENTS v
LIST OF FIGURES vii
LIST OF TABLES viii
PREFACE ix
Chapter 1: Introduction 1
1.1 Introduction 1
1.2 Problem Statement 2
1.3 Research Objectives 5
1.4 Research Contributions 5
1.5 Dissertation Layout 6
References 7
Chapter 2: Literature Review 9
2.1 Transportation Mode Recognition 9
2.2 Quantifying the Effect of Various Features on BSS 10
2.3 Network and Station-Level BSS Prediction 12
2.4 Quality-Of-Service Measurement for BSS Stations 13
2.5 Summary and Conclusions 14
References 15
Chapter 3: Transportation Mode Recognition 19
3.1 Abstract 19
3.2 Introduction 19
3.3 Related Work 21
3.4 Data Set 22
3.4.1 Data Collection 22
3.4.2 Time Domain Features 22
3.4.3 Frequency Domain Features 23
3.5 Methods 25
3.5.1 K-Nearest Neighbor (KNN) 25
3.5.2 Classification and Regression Tree (CART) 26
3.5.3 Support Vector Machines (SVMs) 26
3.5.4 Random Forest (RF) 27
3.6 Proposed Framework 27
3.7 Data Analysis and Results 31
3.7.1 K-Nearest Neighbors Algorithm (KNN) 31
3.7.2 Classification and Regression Tree (CART) 32
3.7.3 Support Vector Machine (SVM) 32
3.7.4 Random Forest (RF) 33
3.7.5 Heterogeneous Framework RF-SVM 35
3.8 Conclusions and Recommendations for Future Work 36
References 37
Chapter 4: Quantifying the Effect of Various Features on BSS 40
4.1 Abstract 40
4.2 Introduction 40
4.3 Related Work 42
4.4 Methods 43
4.4.1 Count Models 43
4.4.2 Poisson Regression Model (PRM) 44
4.4.3 Negative Binomial Regression Model (NBRM) 44
4.4.4 PRM vs NBRM 44
4.4.5 Random Forest (RF) 45

4.4.6 Bayesian Information Criterion (BIC) 46
4.5 Data Set 47
4.6 Data Analysis and Results 48
4.6.1 Problem Definition and Formulation 48
4.6.2 Random Forest and Bayesian Information Criterion 50
4.6.3 Bike Count Modeling for Each Station 55
4.7 Conclusions and Recommendations for Future Work 58
4.8 Acknowledgements 60
References 60
Chapter 5: Network and Station-Level BSS Prediction 62
5.1 Abstract 62
5.2 Introduction 62
5.3 Related work 64
5.4 Methods 65
5.4.1 Random Forest (RF) 65
5.4.2 Least-Squares Boosting (LSBoost) 66
5.4.3 Partial Least-Squares Regression (PLSR) 66
5.5 Dataset 67
5.6 Data analysis and results 68
5.6.1 Univariate Models 68
5.6.2 Station-Level Analysis 69
5.6.3 Multivariate Model 72
5.7 Conclusions and Recommendations for Future Work 74
5.8 Acknowledgment 75
References 75
Chapter 6: Quality-Of-Service Measurement for BSS Stations 78
6.1 Abstract 78
6.2 Introduction 78
6.3 Related Work 80
6.4 Proposed QoS Measurement 81
6.5 Dataset 83
6.6 Analysis and Results 85
6.6.1 Analysis of Variance (ANOVA) 87
6.6.2 Spatial Analysis 88
6.6.3 Optimal Location of New Stations 90
6.7 Conclusions 92
6.8 Acknowledgements 93
References 93
Chapter 7: Conclusions and Future Research 97

LIST OF FIGURES

Figure 1. The narrative structure of the dissertation 3
Figure 2. Importance of features for different pairs of modes. 28
Figure 3. Classification accuracy for KNN in different cases at different neighbors. 31
Figure 4. Classification accuracy for CART in different cases at different pruning levels. 32
Figure 5. Classification accuracy for RF in different cases at different number of trees. 34
Figure 6. Histogram of the bike counts 47
Figure 7. Stations map [3] 48
Figure 8. BIC before (orange dashed line), and after (blue solid line) feature selection process 51
Figure 9 Importance of the predictors (a) after feature selection (b) of the first proposed solution 52
Figure 10. MPE at each station using the proposed method 57
Figure 11. (a) BIC curve, and (b) fitted model for Station 3 58
Figure 12. Stations map. (Source: Google Maps) 67
Figure 13. RF MAE at different prediction horizons and number of trees. 69
Figure 14. LSBoost MAE at different prediction horizons and number of trees. 69
Figure 15. MaxAE at each station using RF. 71
Figure 16. AE and bike availability at Harry Bridges Plaza Station during a selected week 72
Figure 17. MAE using RF at different prediction horizons. 72
Figure 18. Adjacency matrix of the Bay Area BSS network. 73
Figure 19. PLSR, RF, and LSBoost MAE at different prediction horizons. 74
Figure 20. Stations map [30] 84
Figure 21. Conversion of latitude and longitude of each station to UTM system 85
Figure 22. The locations, and the values of the (a) proposed QoS, and (b) traditionally-known QoS measurements. 86
Figure 23. ANOVA test for Tuesdays of February for the 34 stations for (a) traditionally-known QoS, and (b) proposed QoS 88
Figure 24. The empirical variogram for 45° using transformed coordinates 89
Figure 25. Predicted QoS surface for the case study area 91

LIST OF TABLES

Table 1. Symmary of some past studies [15] 10
Table 2. Transportation mode detection applications [2] 20
Table 3. Measurements of time domain features [12]. 23
Table 4. Summary of some past studies [2]. 25
Table 5. Overall classification accuracy for the SVM using time domain, frequency domain, and pooled features. 33
Table 6. Confusion matrix for SVM using pooled features 33
Table 7. Overall classification accuracy for RF using time domain, frequency domain, and pooled features 34
Table 8. Confusion matrix for RF using pooled features 35
Table 9. Overall classification accuracy for RF-SVM using time domain, frequency domain, and pooled features. 35
Table 10. Confusion matrix for RF-SVM using time domain features 36
Table 11. Confusion matrix for RF-SVM using pooled features 36
Table 12. Log-likelihood of Poisson and negative binomial models 50
Table 13. Estimated parameter values for the NB model for bike availability in the network 55
Table 14. Parameters estimation of the Exponential model for Optimal Occupancy and traditionally-known QoS 90

PREFACE

All the co-authors of the chapters introduced in this dissertation are working at Virginia Polytechnic Institute and State University, Blacksburg, VA.

Hesham A. Rakha is a Samuel Reynolds Pritchard Professor of Engineering, Dept. of Civil & Environmental Engineering, a courtesy Professor, Bradley Dept. of Electrical and Computer Engineering at Virginia Tech and the director of the center for sustainable mobility at Virginia Tech Transportation Institute. He was involved in the early stages of concepts formation and contributed to manuscript edits as the primary Advisor and Committee Chair. Rakha provided extensive guidance toward all the chapters of this dissertation and advice on the research.

Leanna L. House is an Associate Professor of Statistics at Virginia Tech. She was listed as a co-author on the papers presented in Chapter 3 and 6 because of her involvement on statistical analysis. She was also involved in some of the concepts formation stages and contributed to some manuscript edits as the Committee Co-Chair.

Mohammed Elhenawy is a postdoctoral researcher at Virginia Tech Transportation Institute. He was listed as a co-author on the paper presented in Chapter 3, 4, 5, and 6 because of his involvement on data analytics and coding. I was responsible for all major areas of concept formation, coding programming and data analysis, as well as manuscript composition.

Mohammed Almannaa is a PhD student at the Dept. of Civil and Environmental Engineering at Virginia Tech. He was listed as a co-author on the paper presented in Chapter 3 because of his involvement in the manuscript composition.

The research presented in this dissertation was partially funded by the Urban Mobility and Equity Center; and the National Science Foundation UrbComp project.

Chapter 1: Introduction

1.1 Introduction

In the next few decades, many traditional cities will be turned into smart cities, which are greener, safer, and faster. This transformation is supported by recent advances in information and communication technology (ICT) in addition to the expected fast spread of the Internet of Things (IoT) and big data analytics. Smart cities will mitigate some of the negative impacts of traditional cities, which consume 75% of the world's resources and energy and produce 80% of the greenhouse gases [1]. According to [2], a *"Smart City"* is intended as an urban environment, which is able to offer advanced and innovative services to citizens in order to improve the overall quality of their life. Smart cities have many components, including smart transportation. Smart transportation will integrate different transportation networks and allow them to work together so travelers and commuters can enjoy seamless multi-mode trips based on their preferences. Consequently, more commuters will be inspired to use public transportation systems and many traffic-related problems such as congestion will be relaxed.

The last mile problem is a pressing problem that needs to be solved in order for different transportation networks to work together efficiently. This problem is defined as *"the short distance between home and public transit or transit stations and the workplace, which may be too far to walk. [3]"* One solution to this problem is a bike-sharing system (BSS), which takes advantage of the availability of ICT and the BSS's data to smartly operate the network. Smart bike-sharing systems (SBSSs) use recent technologies to monitor the status of each station in the network, collect bike usage data and other relevant data, and use state-of-the-art algorithms to build predictive models, predict future bike availability, and find good solutions for the issue of imbalance in the distribution of bikes in order to guarantee users' satisfaction and meet their demand.

Due to relatively low capital and operational costs, as well as ease of installation, many cities in the U.S. are making investments in BSSs. A technical report distributed by the Bureau of Transportation in April 2016 indicated that there are 2,655 BSSs stations in 65 U.S. cities, and that 86.3% of these stations are connected to another means of scheduled public transportation [4]. These numbers show that the physical infrastructure for BSSs already exists and that they are good

candidates for connecting different transportation networks. In 2013, San Francisco launched the Bay Area Bike Share System (BSS) (now called the "Ford GoBike" BSS), a membership-based system providing 24-hours-per-day, 7-days-per-week self-service access to short-term rental bicycles. Members can check out a bicycle from a network of automated stations, ride to the station nearest their destination, and leave the bicycle safely locked for someone else to use [5]. The Bay Area Bike Share is designed for short, quick trips, and as a result, additional fees apply for trips longer than 30 minutes. In this system, 70 bike stations connect users to transit, businesses, and other destinations in four areas: downtown San Francisco, Palo Alto, Mountain View, and downtown San Jose [5]. Bay Area Bike Share is available to everyone 18 years and older with a credit or debit card. The system is designed to be used by commuters and tourists alike, whether they are trying to get across town at rush hour, traveling to and from Bay Area Rapid Transit (BART) and Caltrain stations, or pursuing daily activities [5].

Bike sharing is a public system in which customers can automatically lend bikes at one station and delivered at another within a restricted time period. The definition of a well-function BSS is the system where there are bikes available at a station when someone wants to pick up a bike and that there are free docks available when someone wants to return one. In order to achieve this, most operator use service vehicles to rebalance the system, i.e. they move bikes from (nearly) full stations to (nearly) empty stations [6].

1.2 Problem Statement

BSSs suffer from several planning problem that could be divided into three levels [6, 7]; a strategic, a tactical, and an operational level. Strategic level includes problem in the stage of designing BSSs such as investigating the effect of different factors on the system and determining the optimal number of bikes and location of stations. On the tactical level, the focus is on finding an optimal distribution of bikes between the stations at a specific time, i.e. rebalancing. Operational level problems arise from finding optimal routes for service vehicles to rebalance the system.

This dissertation is a building block for a smart BSS at the strategic level, which could be used in real different applications. Generally, four components will be developed in the dissertation: transportation mode recognition, quantifying the effect of various features on BSS, network and station-level BSS prediction, and quality-of-service measurement for BSS stations. Figure 1 shows the narrative structure of the dissertation. At the strategic level, we start with

proposing a novel two-layer hierarchical classifier that increases the accuracy of traditional transportation mode classification algorithms, which can be utilized in planning new BSS instead of using costly surveys. However, the other three components of the dissertation were developed to be mainly applied for an existing BSS. A new method is proposed to quantify the effect of several factors such as weather conditions on the prediction of bike counts at each station. For the third components, these resulted significant features were used to develop state-of-the-art toolbox algorithms to operate BSSs efficiently at two levels: network and station. The fourth components used the toolbox algorithms to estimate the proposed QoS measurement.

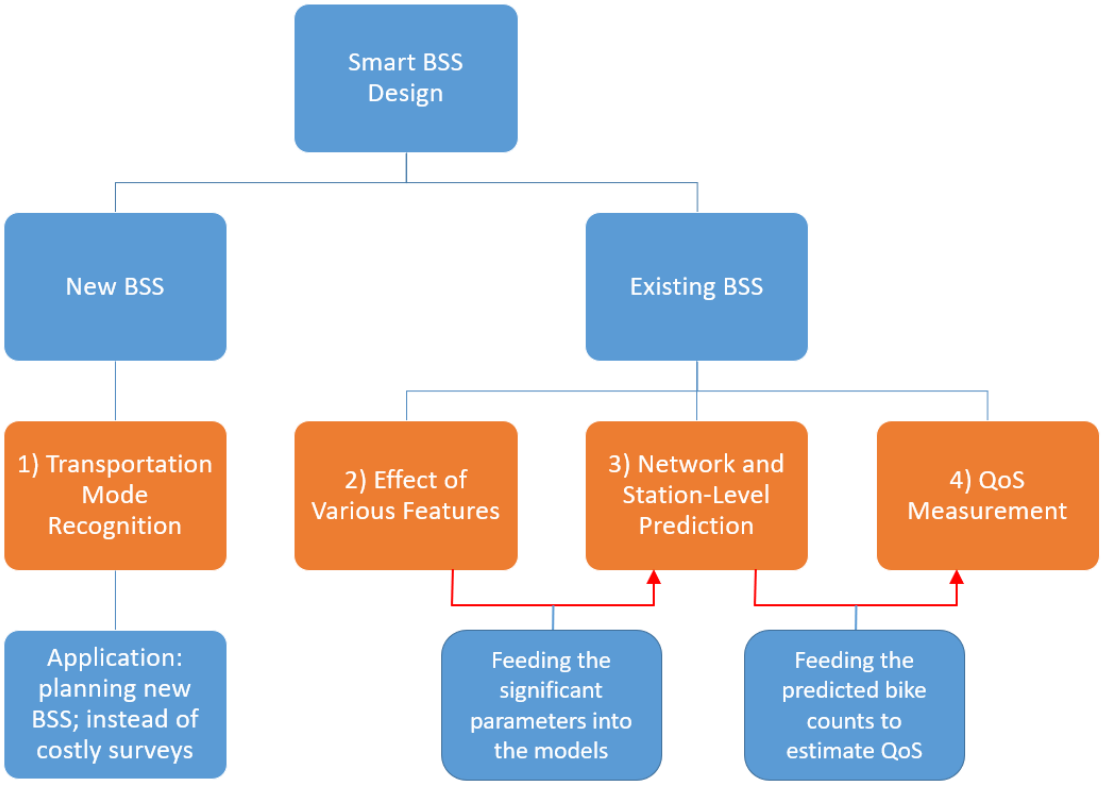

Figure 1. The narrative structure of the dissertation

In the first component, we investigated the possibility of improving the overall accuracy of transportation mode detection by proposing a new hierarchical framework classifier and by looking for a new feature set. Most of the proposed methods in the most recent studies rely on the using of the GPS data, which do not consider the limitations of GPS information. GPS service is not available or may be lost in some areas, which results in inaccurate position information. Moreover, the GPS system use might deplete the smartphone's battery. Thus, this paper focuses on proposing a new detection framework using machine learning techniques and extract new

features based on data obtained from smartphone sensors including accelerometer, gyroscope, and rotation vector, without GPS data.

In the second component, a novel approach was proposed to construct a bike count model for the San Francisco Bay Area BSS. The bike counts at each station, each of which has a finite number of docks, fluctuates. Thus, a rebalancing (or redistribution) operation must be performed periodically to meet this fluctuation. Coordinating such a large operation is complicated, time consuming, polluting and expensive [8]. Firstly, we quantify the effect of several variables on the mean of bike counts for the Bay Area BSS network, including the month-of-the-year, the day-of-the-week, time-of-the-day, and various weather conditions. Secondly, using the same proposed method, the paper constructs a predictive model for the bike counts at each station over the time as it is one of the key tasks to making the rebalancing operation more efficient.

The third component develops models for modeling and predicting the availability of bikes in the San Francisco Bay Area Bike Share System using machine learning algorithms at two levels: network and station. Bike count prediction at the station-level using machine learning algorithms has not been studied well to date. While the full prediction problem would be predicting bike counts at each station, the previous studies used machine learning to predict the bike count of the entire BSS instead (to be used for clustering for example). However, the state-of-the-art models that were used to predict the number of available bikes at each station ignore the correlation between stations and might become hard to implement when applied to relatively large networks. Moreover, we investigate the use of multivariate response models to predict the number of available bikes in the network, which has not been addressed before.

As for the fourth component, we propose a new discriminative QoS measurement that reflects the spatial dependencies in an inhomogeneous BSS and that considers the variability of arrival and pickup rates. Then, we use this QoS measurement with geo-statistics to model a spatial variogram that could predict the QoS in nearby areas for the purpose of locating new stations in an existing BSS. The traditionally-known QoS measurement in the literature is based on the proportion of *problematic stations*, which are defined as those with no bikes or docks available for users. We thoroughly investigated the traditionally-known QoS measurement, and it was found neither satisfying in exposing the spatial dependencies between stations nor discriminative in describing different stations in a BSS. BSS operators take great efforts to ensure bike and dock availability in each station. This task can be difficult as the movements of users are highly dynamic,

difficult to predict, and redistributing bikes is expensive. Recent studies have shown that there are spatial dependencies in bike usage at different stations [9-12], and that imbalances in the spatial distribution of bikes occur due to one-way use and short rental periods [12]. Thus, it is necessary for operators to understand the spatial dependencies to more effectively manage the system. For example, operators could improve the quality of service (QoS) by identifying the best candidate spots for new stations.

1.3 Research Objectives

According to the above discussion and in light of the mentioned limitations, this research effort attempt to develop some of the components toward strategically designing a smart BSS. To achieve this goal several objectives are addressed. We start with proposing a novel two-layer hierarchical classifier that increases the accuracy of traditional transportation mode classification algorithms. In the transportation sector, researchers can use smartphones to track and obtain information of multi-mode trips. These data can be used to recognize the user's transportation mode, which can be then utilized in a number of different applications.

Next, a new method is proposed to quantify the effect of several factors such as weather conditions on the prediction of bike station counts. The proposed approach is promising to quantify the effect of various features on BSSs in cases of large networks with big data. Third, we use state-of-the-art data analytics to develop a toolbox to operate BSSs efficiently at two levels: network and station. Fourth, we propose a quality-of-service (QoS) measurement, namely *Optimal Occupancy*, which considers the impact of inhomogeneity in a BSS.

In general, these objectives could be used to boost the redistribution operation [13-15], to gain new insights into and correlations between bike demand and other factors [16-19], and to support policy makers and managers in making good decisions [12, 16].

1.4 Research Contributions

This dissertation makes many significant contributions to the literature. These contributions include novel methods, measurements, and applications using machine learning and statistical learning techniques in order to design a smart BSS in the strategic level. Specifically, the contributions of this research effort can be summarized as follows:

- We proposed a two-layer hierarchical framework in which a) the first layer contains one multi-classifier using the data set of the five transportation modes, and b) the second layer consists of 10 binary classifiers, each of which is specialized in only one pair of modes and uses a features subset that discriminates between this pair.
- A new frequency domain features were extracted and pooled with the traditionally-used time domain features to improve the overall accuracy of transportation mode detection.
- We introduced an effective and fast, but also accurate and reasonable, approach to quantifying the effect of various features on bike counts at different stations. This method is promising in quantifying the effect of various features in cases of large networks with big data.
- We modeled the bike availability at the station-level using machine learning algorithms, which has not been studied well to date.
- The univariate response models previously used to predict the number of available bikes at each station ignore the correlation between stations and might become hard to implement when applied to relatively large networks. We investigated the use of multivariate response models to reduce the number of required prediction models and reflect the spatial correlation between stations at the network-level.
- Station neighbors, which are determined by a trip's adjacency matrix, are considered as significant predictors in the regression models.
- We proposed a new discriminative QoS measurement that reflects the spatial dependencies in an inhomogeneous BSS and that considers the variability of arrival and pickup rates.
- We used the new QoS measurement with geo-statistics to model a spatial variogram that could predict the QoS in nearby areas for the purpose of locating new stations in an existing BSS.

1.5 Dissertation Layout

After the introduction chapter, which describes the problem statement and dissertation objectives, an extensive review of the literature relevant to topics covered in the dissertation is presented in Chapter 2. Thereafter, Chapter 3 includes the mode transportation recognition component, which describes the data and methodology used to develop the proposed framework. Chapter 4 provides the detailed method used in quantifying the effect of various features on BSS.

Chapter 5 develops models for modeling and predicting the availability of bikes in the San Francisco Bay Area Bike Share System using machine learning algorithms at two levels: network and station. Chapter 6 proposes a new discriminative QoS measurement that reflects the spatial dependencies in an inhomogeneous BSS. Finally, the seventh chapter consists of a summary of the dissertation conclusions and recommendations for further research.

References

[1] S. P. Mohanty, U. Choppali, and E. Kougianos, "Everything you wanted to know about smart cities: The internet of things is the backbone," *IEEE Consumer Electronics Magazine*, vol. 5, no. 3, pp. 60-70, 2016.

[2] G. Piro, I. Cianci, L. A. Grieco, G. Boggia, and P. Camarda, "Information centric services in smart cities," *Journal of Systems and Software*, vol. 88, pp. 169-188, 2014.

[3] S. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in Europe, the Americas, and Asia: past, present, and future," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2143, pp. 159-167, 2010.

[4] Bureau of Transportation Statistics. (2018). *National Transportation Statistics*. Available: https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/index.html

[5] Bay Area Bike Share. (2016). *Introducing Bay Area Bike Share, your new regional transit system*. Available: http://www.bayareabikeshare.com/faq#BikeShare101

[6] H. M. Espegren, J. Kristianslund, H. Andersson, and K. Fagerholt, "The Static Bicycle Repositioning Problem-Literature Survey and New Formulation," 2016, pp. 337-351: Springer.

[7] P. Vogel, "Service network design of bike sharing systems," in *Service Network Design of Bike Sharing Systems*: Springer, 2016, pp. 113-135.

[8] P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, p. 3, 2009.

[9] P. Borgnat, E. Fleury, C. Robardet, and A. Scherrer, "Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program," in *ECCS'09*, Warwick, United Kingdom, 2009: Complex Systems Society.

[10] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," 2009, vol. 9, pp. 1420-1426.

[11] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455-466, 2010.

[12] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns," *Procedia - Social and Behavioral Sciences*, vol. 20, pp. 514-523, 2011/01/01 2011.

[13] C. Contardo, C. Morency, and L.-M. Rousseau, *Balancing a dynamic public bike-sharing system*. Cirrelt, 2012.

[14] J. Schuijbroek, R. Hampshire, and W.-J. van Hoeve, "Inventory rebalancing and vehicle routing in bike sharing systems," 2013.

[15] T. Raviv, M. Tzur, and I. A. Forma, "Static repositioning in a bike-sharing system: models and solution approaches," *EURO Journal on Transportation and Logistics,* vol. 2, no. 3, pp. 187-229, 2013// 2013.
[16] D. W. Daddio, "Maximizing Bicycle Sharing: an empirical analysis of capital bikeshare usage," 2012.
[17] X. Wang, G. Lindsey, J. E. Schoner, and A. Harrison, "Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations," *Journal of Urban Planning and Development,* vol. 142, no. 1, p. 04015001, 2015.
[18] C. Rudloff and B. Lackner, "Modeling demand for bicycle sharing systems–neighboring stations as a source for demand and a reason for structural breaks," 2013.
[19] R. Rixey, "Station-Level Forecasting of Bikesharing Ridership: Station Network Effects in Three US Systems," *Transportation Research Record: Journal of the Transportation Research Board,* no. 2387, pp. 46-55, 2013.

Chapter 2: Literature Review

To further determine the significant of this dissertation and the proposed methods and measurements, this chapter is divided into four sections that covers the research objectives: 1) Transportation mode recognition, 2) Quantifying the effect of various features on BSS, 3) Network and station-level BSS prediction, and 4) Quality-of-service measurement for BSS stations.

2.1 Transportation Mode Recognition

Researchers have developed several approaches to discriminate between transportation modes effectively using mobile phones [1, 2] or visual tracking [3]. Machine learning techniques have been used extensively to build detection models and have shown high accuracy in determining transportation modes. Supervised learning methods such as K-Nearest Neighbor (KNN) [4], Support Vector Machines (SVMs) [5-10], Decision Trees [6, 7, 11-14], and Random Forests (RFs) [4], have all been employed in various studies.

These studies have obtained different classifying accuracies. There are several factors that affect the accuracy of detecting transportation modes, such as the *monitoring period* (positive association), *number of modes* (negative association), *data sources, motorized classes,* and *sensor positioning* [4, 15].

However, one of the most critical factors that affects the accuracy of mode detection is the machine learning framework classifier. The framework that usually uses one layer of classification algorithm as in [6, 16] could be refered to as traditional framework; whereas the hierarchal framework uses more than one layer of classification algorithm.

An additional important consideration is the domain of the extracted features. Features are generally extracted from two different domains: (1) the time domain, features of which have been used widely in many studies [4, 8, 9, 13, 17, 18]; and (2) the frequency domain, features of which have been used in some studies [6, 8]. Both methods have achieved a significant, high accuracy. Table 1 summarizes the obtained accuracies and factors for some of the aforementioned studies. Note that no direct comparison can be made between the studies listed in Table 1 because the factors considered, and the data sets used varied from study to study.

In this study, we will mainly focus in the effect of two factors on the accuracy of transportation mode detection; namely the framework classifier and extracted features. Most of the

proposed methods in the most recent studies rely on the using of the GPS data, which do not consider the limitations of GPS information. GPS service is not available or may be lost in some areas, which results in inaccurate position information. Moreover, the GPS system use might deplete the smartphone's battery. Thus, this paper focuses on proposing a new detection framework using machine learning techniques and extract new features based on data obtained from smartphone sensors including accelerometer, gyroscope, and rotation vector, without GPS data.

Table 1. Symmary of some past studies [15]

Accuracy (%)	Features Domain	Machine Learning Framework	Monitoring Period	No. of Modes	Data Sources	More than One Motorized Mode?	Sensor Positioning	Data Set	Study
97.31	Time	Traditional	4 s	3	Accelerometer	Yes	No requirements	Not mentioned	[9]
93.88	Frequency	Traditional	5 s, 50% overlap	6	Accelerometer	Yes/No	Participants were asked to keep their device in the pocket of their non-dominant hip	Collected from 4 participants	[8]
93.60	Time and frequency	Traditional	1 s	5	Accelerometer GPS	No	No requirements	Collected from 16 participants	[6]
93.50	Time	Traditional	30 s	6	GPS, GIS[a] maps	Yes	No requirements	Collected from 6 participants	[16]
95.10	Time	Traditional	1 s	5	Accelerometer, gyroscope, rotation vector	Yes	No requirements	Collected from 10 participants	[4]
91.60	Time	Traditional	Entire trip	11	GPS, GIS maps	Yes	No requirements	Two different data sets, one of which included 1,000 participants	[18]

[a] GIS: Geographic Information System

2.2 Quantifying the Effect of Various Features on BSS

The modeling of BSS data using various features, including time, weather, built-environment, transportation infrastructure, etc., is an area of significant research interest. In general, the main goals of data modeling are to boost the redistribution operation [19-21], to gain new insights into and correlations between bike demand and other factors [22-25], and to support policy makers and managers in making good decisions [22, 26]. Generally, the main approach to modeling and predicting bike sharing data is regression count modeling. A recent paper modeled the demand for bikes and return boxes using data from the BSS Citybike Wien in Vienna, Austria. The influence of weather (temperature and precipitation) and full/empty neighboring stations on demand was studied using different count models (Poisson, Negative Binomial [NB] and Hurdle).

The authors found that although the Hurdle model worked best in modeling the demand of bike sharing stations, these models were complex and might not be ideal for optimization procedures. They also found that NB models outperformed Poisson models because of the dispersion in the data (to be discussed later) [24]. However, an early study used count series to predict the stations' usage based on Poisson mixtures, providing insight into the relationship between station neighborhood type and mobility patterns [27].

In a study by Wang et al., log-linear and NB regression models were used to estimate total station activity counts. The factors used included: economical, built-environment, transportation infrastructural and social aspects, such as neighborhood sociodemographic (i.e., age and race), proximity to the central business district, proximity to water, accessibility to trails, distance to other bike share stations, and measures of economic activity. All the variables were found to be significant. The Log-likelihood was used as a measure of the goodness of fit of the Poisson and NB models [23]. Linear least squared regression with data from the on-the-ground Capital BSS was implemented in another paper to predict station demand based on demographic, socioeconomic, and built-environment characteristics [22].

Several studies used methods other than count models to model BSS data. A multivariate linear regression analysis was used in another study to study station-level BSS ridership. That study investigated the correlation between BSS ridership and the following factors: population density; retail job density; bike, walk, and transit commuters; median income; education; presence of bikeways; nonwhite population (negative association); days of precipitation (negative association); and proximity to a network of other BSS stations. The authors found that the demographic, built environment, and access to a comprehensive network of stations were critical factors in supporting ridership [25].

A study by Gallop et al. used continuous and year-round hourly bicycle counts and weather data to model bicycle traffic in Vancouver, Canada. The study used seasonal autoregressive integrated moving average analysis to account for the complex serial correlation patterns in the error terms and tested the model against actual bicycle traffic counts. The study demonstrated that the weather had a significant and important impact on bike usage. The authors found that the weather data (namely temperature, rain, and clearness) were generally significant; temperature and rain, specifically, had an important effect [28].

It is also worth noting that some studies used methods other than regression to either model BSS data or to develop new insights and understandings of BSSs (see [19, 26]). For example, a mathematical formulation for the dynamic public bike-sharing balancing problem was introduced using two different models: the arc-flow formulation and the Dantzig-Wolfe decomposition formulation. The demand was computed by considering the station either a pickup or delivery point, with a real-time and length period between two stations [19].

2.3 Network and Station-Level BSS Prediction

Modeling bike sharing data is an area of significant research interest. Proposed models have relied on various features, including time, weather, the built environment, and transportation infrastructure. Many studies have been performed at the station-level to predict the availability of bikes by using time series analysis. Rixey used a multivariate linear regression analysis to study station-level BSS ridership, investigating the correlation between BSS ridership and the following factors: population density; retail job density; bike, walk, and transit commuters; median income; education; presence of bikeways; nonwhite population (negative association); days of precipitation (negative association); and proximity to a network of other BSS stations.[25] The author found that demographics, the built environment, and access to a comprehensive network of stations were critical factors in supporting ridership.

Froehlich, Neumann, and Oliver used four predictive models to predict the number of available bikes at each station: last value, historical mean, historical trend, and Bayesian networks [29]. Two methods for time series analysis, autoregressive-moving average (ARMA) and autoregressive integrated moving average (ARIMA), have also been used to predict the number of available bikes/docks at each bike station. Kaltenbrunner et al. adopted ARMA [30]; Yoon et al. proposed a modified ARIMA model considering spatial interaction and temporal factors [31]. However, Gallop et al. used continuous and year-round hourly bicycle counts and weather data to model bicycle traffic in Vancouver, Canada [28]. That study used a seasonal autoregressive integrated moving average analysis to account for the complex serial correlation patterns in the error terms and tested the model against actual bicycle traffic counts. The results demonstrated that the weather had a significant and important impact on bike usage. The authors found that the weather data (i.e., temperature, rain, humidity, and clearness) were generally significant and that temperature and rain, specifically, had an important effect.

However, few studies have used machine learning to model bike sharing data. One of the characteristics of transportation-related datasets is that they are often very large. It is therefore advantageous to implement machine learning to identify potential explanatory variables [26]. Moreover, when a model contains a large number of predictors, it becomes more complex and overfitting can occur. To address this, different algorithms have been used to predict bike availability in a BSS, such as random forest (RF), support vector machine (SVM), and gradient boosted tree (GBT) [32-37]. The authors of the four studies in [32, 33, 35, 36] used different machine learning algorithms to predict bike demand based on the usage record and other information about the targeting prediction time window. While the full prediction problem would be predicting bike counts at each station, the authors used machine learning to predict the bike count of the entire BSS instead. In [34], the authors used RF to classify the status of the stations only with regard to whether the station was completely full of bikes or completely empty, so users could not return a bicycle, or could not find one to rent.

2.4 Quality-Of-Service Measurement for BSS Stations

Modeling bike sharing data is an area of significant research interest. Research questions that have been studied previously include the strategic design, operation, and analysis of BSSs. Due to the potential benefits to operators, measuring the service level of stations or the whole system [38] has become an appealing issue for researchers. In some cases, operators measure the fraction of time that their stations are full or empty as a measurement of the QoS of the system [20]. Similarly, Fricker *et al.* considered the limiting probability that a station is empty or full as the performance measure. They argued that the optimal proportion of bikes per station is slightly more than half the capacity of a station in a homogeneous system. In an inhomogeneous system, however, they concluded that this performance metric collapses due to the heterogeneity [39].

Lin and Yang [40] investigated the strategic problems by studying the question of bike stations' measures of service. They argued that the measures of service quality in the system should include two measurements: the availability rate, which was defined as the proportion of pick-up requests at a bike station that are met by the bicycle stock on hand, and the coverage level, which is the fraction of total demand at both origins and destinations that is within some specified time or distance from the nearest rental station. Fricker and Gast [41] proposed a stochastic model of a homogeneous BSS and investigated the impact of users' random choices on the number of

problematic stations. Problematic stations were defined as stations that, at a given time, have no bikes available or no available spots for bikes to be returned to. Therefore, the performance of the system was determined by the proportion of problematic stations. However, these measures have critical drawbacks: (1) as BSSs usually offer two services: picking up bikes, and returning bikes; these measurements fail to take into account the service quality of returning bikes to stations; (2) some of the studies assume that, in contrast to real systems, the system is homogeneous; and (3) while some studies modeled the system as inhomogeneous, they failed to consider the variability of the system parameters (i.e., arrival and pickup rates) throughout the same day or across the different days of the week and their dependency on the individual station.

In any BSS, one of the keys to success is the location and distribution of bike stations [40]. Some studies have worked on locating bike stations using different methods, such as location-allocation models [42, 43], and an optimization method that maximizes the demand covered and takes the available budget as a constraint [44]. The spatial distribution of the *potential demand* is a fundamental element in optimal location modeling. In order to estimate potential demand, several studies used preference surveys to evaluate both the factors influencing the use of the bicycle mode and choice of routing [45-48]. Potential demand has also been estimated by considering the population, employment associated with each building number, and the number of trips generated for each transport zone [42]. However, there are some limitations and drawbacks in the methods previously used to find the optimal station location: these methods are basically used to plan new systems and might not be useful to predict new stations in existing systems; they are aimed at serving the local population on selected days (e.g., workdays); and certain places in the studied area (e.g., large parks) have neither population nor jobs and yet may attract a considerable number of trips.

2.5 Summary and Conclusions

This chapter provides an extensive literature review to the four components that covers the research objectives. The first component proposed a new detection framework using machine learning techniques and extract new features based on data obtained from smartphone sensors including accelerometer, gyroscope, and rotation vector, without GPS data (due to its limitations). In the second component, we found that previous studies have chosen specific weather variables depending on expert judgment and other state-of-the-art studies without really investigating their

significance. In this study, we propose a method that will allow the operators to quantify the significance of various features in cases of large networks of big data. Nonetheless, we studied the full prediction problem in the third component, we found that previous studies used machine learning to predict the bike count of the entire BSS only. Thus, we use state-of-the-art data analytics to develop a toolbox to operate BSSs efficiently at the two levels: network and station. In the fourth component, the traditionally-known QoS measurement in the literature was thoroughly investigated, and it was found neither satisfying in exposing the spatial dependencies between stations nor discriminative in describing different stations in a BSS. As a result, we propose a new QoS measurement to overcome the abovementioned issues in inhomogeneous BSSs.

References

[1] M. Susi, V. Renaudin, and G. Lachapelle, "Motion mode recognition and step detection algorithms for mobile phone users," *Sensors*, vol. 13, no. 2, pp. 1539-1562, 2013.

[2] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74-82, 2011.

[3] L. Wang, L. Zhang, and Z. Yi, "Trajectory predictor by using recurrent neural networks in visual tracking," *IEEE transactions on cybernetics*, vol. 47, no. 10, pp. 3172-3183, 2017.

[4] A. Jahangiri and H. A. Rakha, "Applying machine learning techniques to transportation mode recognition using mobile phone sensor data," *IEEE transactions on intelligent transportation systems*, vol. 16, no. 5, pp. 2406-2417, 2015.

[5] L. Zhang, M. Qiang, and G. Yang, "Mobility transportation mode detection based on trajectory segment," *Journal of Computational Information Systems*, vol. 9, no. 8, pp. 3279-3286, 2013.

[6] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using Mobile Phones to Determine Transportation Modes," *Acm Transactions on Sensor Networks*, vol. 6, no. 2, Feb 2010, Art. no. 13.

[7] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the web," presented at the Proceedings of the 17th international conference on World Wide Web, Beijing, China, 2008.

[8] B. Nham, K. Siangliulue, and S. Yeung, "Predicting mode of transport from iphone accelerometer data," Tech. report, Stanford Univ2008.

[9] T. Nick, E. Coersmeier, J. Geldmacher, and J. Goetze, "Classifying means of transportation using mobile sensor data," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, 2010, pp. 1-6: IEEE.

[10] A. Bolbol, T. Cheng, I. Tsapakis, and J. Haworth, "Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification," *Computers, Environment and Urban Systems*, 2012.

[11] X. Yu *et al.*, "Transportation activity analysis using smartphones," in *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, 2012, pp. 60-61.

[12] P. Widhalm, P. Nitsche, and N. Brandie, "Transport mode detection with realistic Smartphone sensor data," in *2012 21st International Conference on Pattern Recognition (ICPR 2012), 11-15 Nov. 2012*, Piscataway, NJ, USA, 2012, pp. 573-6: IEEE.

[13] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu, "Transportation mode detection using mobile phones and GIS information," in *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS 2011, November 1, 2011 - November 4, 2011*, Chicago, IL, United states, 2011, pp. 54-63: Association for Computing Machinery.

[14] V. Manzoni, D. Maniloff, K. Kloeckl, and C. Ratti, "Transportation mode identification and real-time CO2 emission estimation using smartphones," Technical report, Massachusetts Institute of Technology, Cambridge2010.

[15] M. Elhenawy, A. Jahangiri, and H. A. Rakha, "Smartphone Transportation Mode Recognition using a Hierarchical Machine Learning Classifier," presented at the 23rd ITS World Congress, MELBOURNE AUSTRALIA, October 2016, 2016. Available: https://www.researchgate.net/publication/301338082

[16] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu, "Transportation mode detection using mobile phones and GIS information," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2011, pp. 54-63: ACM.

[17] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, no. 2, p. 13, 2010.

[18] F. Biljecki, H. Ledoux, and P. Van Oosterom, "Transportation mode-based segmentation and classification of movement trajectories," *International Journal of Geographical Information Science*, vol. 27, no. 2, pp. 385-407, 2013.

[19] C. Contardo, C. Morency, and L.-M. Rousseau, *Balancing a dynamic public bike-sharing system*. Cirrelt, 2012.

[20] J. Schuijbroek, R. Hampshire, and W.-J. van Hoeve, "Inventory rebalancing and vehicle routing in bike sharing systems," 2013.

[21] T. Raviv, M. Tzur, and I. A. Forma, "Static repositioning in a bike-sharing system: models and solution approaches," *EURO Journal on Transportation and Logistics*, vol. 2, no. 3, pp. 187-229, 2013// 2013.

[22] D. W. Daddio, "Maximizing Bicycle Sharing: an empirical analysis of capital bikeshare usage," 2012.

[23] X. Wang, G. Lindsey, J. E. Schoner, and A. Harrison, "Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations," *Journal of Urban Planning and Development*, vol. 142, no. 1, p. 04015001, 2015.

[24] C. Rudloff and B. Lackner, "Modeling demand for bicycle sharing systems–neighboring stations as a source for demand and a reason for structural breaks," 2013.

[25] R. Rixey, "Station-Level Forecasting of Bikesharing Ridership: Station Network Effects in Three US Systems," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2387, pp. 46-55, 2013.

[26] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns," *Procedia - Social and Behavioral Sciences*, vol. 20, pp. 514-523, 2011/01/01 2011.

[27] C. Etienne and O. Latifa, "Model-Based Count Series Clustering for Bike Sharing System Usage Mining: A Case Study with the Velib' System of Paris," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1-21, 2014.
[28] C. Gallop, C. Tse, and J. Zhao, "A seasonal autoregressive model of Vancouver bicycle traffic using weather variables," *i-Manager's Journal on Civil Engineering*, vol. 1, no. 4, p. 9, 2011.
[29] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," 2009, vol. 9, pp. 1420-1426.
[30] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455-466, 2010.
[31] J. W. Yoon, F. Pinelli, and F. Calabrese, "Cityride: a predictive bike sharing journey advisor," 2012, pp. 306-311: IEEE.
[32] Y.-C. Yin, C.-S. Lee, and Y.-P. Wong, "Demand Prediction of Bicycle Sharing Systems," ed: Stanford University.[Online], 2012.
[33] J. Du, R. He, and Z. Zhechev, "Forecasting Bike Rental Demand," ed: Stanford University, 2014.
[34] G. M. Dias, B. Bellalta, and S. Oechsner, "Predicting occupancy trends in Barcelona's bicycle service stations using open data," 2015, pp. 439-445: IEEE.
[35] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," 2015, p. 33: ACM.
[36] R. Giot and R. Cherrier, "Predicting bikeshare system usage up to one day ahead," 2014, pp. 22-29: IEEE.
[37] H. I. Ashqar, M. Elhenawy, M. H. Almannaa, A. Ghanem, H. A. Rakha, and L. House, "Modeling bike availability in a bike-sharing system using machine learning," in *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2017, pp. 374-378.
[38] A. Gunasekaran, C. Patel, and E. Tirtiroglu, "Performance measures and metrics in a supply chain environment," *International journal of operations & production Management*, vol. 21, no. 1/2, pp. 71-87, 2001.
[39] C. Fricker, N. Gast, and H. Mohamed, "Mean field analysis for inhomogeneous bike sharing systems," 2012.
[40] J.-R. Lin and T.-H. Yang, "Strategic design of public bicycle sharing systems with service level constraints," *Transportation research part E: logistics and transportation review*, vol. 47, no. 2, pp. 284-294, 2011.
[41] C. Fricker and N. Gast, "Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity," *Euro journal on transportation and logistics*, vol. 5, no. 3, pp. 261-291, 2016.
[42] J. C. García-Palomares, J. Gutiérrez, and M. Latorre, "Optimizing the location of stations in bike-sharing programs: A GIS approach," *Applied Geography*, vol. 35, no. 1, pp. 235-246, 2012/11/01/ 2012.
[43] G. Rybarczyk and C. Wu, "Bicycle facility planning using GIS and multi-criteria decision analysis," *Applied Geography*, vol. 30, no. 2, pp. 282-293, 2010.
[44] I. Frade and A. Ribeiro, "Bike-sharing stations: A maximal covering location approach," *Transportation Research Part A: Policy and Practice*, vol. 82, no. Supplement C, pp. 216-227, 2015/12/01/ 2015.

[45] J. Dill and K. Voros, "Factors affecting bicycling demand: initial survey findings from the Portland, Oregon, region," *Transportation Research Record: Journal of the Transportation Research Board,* no. 2031, pp. 9-17, 2007.
[46] J. E. Abraham, S. McMillan, A. T. Brownlee, and J. D. Hunt, "Investigation of cycling sensitivities," 2002.
[47] K. Shafizadeh and D. Niemeier, "Bicycle journey-to-work: travel behavior characteristics and spatial attributes," *Transportation Research Record: Journal of the Transportation Research Board,* no. 1578, pp. 84-90, 1997.
[48] L. D. O. Meng, "Implementing bike-sharing systems," *Proceedings of the Institution of Civil Engineers,* vol. 164, no. 2, p. 89, 2011.

Chapter 3: Transportation Mode Recognition

This chapter based on

Ashqar, Huthaifa I., Mohammed H. Almannaa, Mohammed Elhenawy, Hesham A. Rakha, and Leanna House. "Smartphone Transportation Mode Recognition Using a Hierarchical Machine Learning Classifier and Pooled Features from Time and Frequency Domains." *IEEE Transactions on Intelligent Transportation Systems* (2018).

3.1 Abstract

The paper develops a novel two-layer hierarchical classifier that increases the accuracy of traditional transportation mode classification algorithms. The study also enhances classification accuracy by extracting new frequency domain features. Many researchers have obtained these features from Global Positioning System (GPS) data; however, this data was excluded in our study, as the system use might deplete the smartphone's battery and signals may be lost in some areas. Our proposed two-layer framework differs from previous classification attempts in three distinct ways: (1) the outputs of the two layers are combined using Bayes' rule to choose the transportation mode with the largest posterior probability; (2) the proposed framework combines the new extracted features with traditionally used time domain features to create a pool of features; (3) a different subset of extracted features is used in each layer based on the classified modes. Several machine learning techniques were used, including k-nearest neighbor, classification and regression tree, support vector machine, random forest, and a heterogeneous framework of random forest and support vector machine. Results show that the classification accuracy of the proposed framework outperforms traditional approaches. Transforming the time domain features to the frequency domain also adds new features in a new space and provides more control on the loss of information. Consequently, combining the time domain and the frequency domain features in a large pool and then choosing the best subset results in higher accuracy than using either domain alone. The proposed two-layer classifier obtained a maximum classification accuracy of 97.02%.

3.2 Introduction

The application of smartphones to data collection has recently attracted researchers' attention. Smartphone applications (apps) have been developed and effectively used to collect data

from smartphones in many sectors. In the transportation sector, researchers can use smartphones to track and obtain information such as speed, acceleration, and the rotation vector from the built-in Global Positioning System (GPS), accelerometer, and gyroscope sensors [1]. These data can be used to recognize the user's transportation mode, which can be then be utilized in a number of different applications, as shown in Table 2.

Table 2. Transportation mode detection applications [2]

Application	Description
Transportation Planning	Instead of using traditional approaches such as questionnaires, travel diaries, and telephone interviews [3, 4], the transportation mode information can be automatically obtained through smartphone sensors.
Safety	Knowing the transportation mode can help in developing safety applications. For example, violation prediction models have been studied for passenger cars and bicycles [5, 6].
Environment	Physical activities, health, and calories burned, and carbon footprint associated with each transportation mode can be obtained when the mode information is available [7].
Information Provision	Traveler information can be provided based on the transportation mode [4, 8].

In this study, we investigated the possibility of improving the overall accuracy of transportation mode detection by proposing a new hierarchical framework classifier and by looking for a new feature set. This paper makes two major contributions to the body of transportation research. First, it proposes a two-layer hierarchical framework in which a) the first layer contains one multi-classifier using the data set of the five transportation modes, and b) the second layer consists of 10 binary classifiers, each of which is specialized in only one pair of modes and uses a features subset that discriminates between this pair. Second, new frequency domain features were extracted and pooled with the traditionally-used time domain features.

Following the introduction, this paper is organized into six sections. First, the approaches, features, and machine learning techniques of previous studies are reviewed. Next, the data set and the extracted features are described. Third, background is presented on the machine learning techniques applied in this study. Next, the proposed framework is presented. In the fifth section,

details are provided on the data analysis used to detect different transportation modes. Finally, the paper concludes with a summary of new insights and recommendations for future transportation mode recognition research.

3.3 Related Work

Researchers have developed several approaches to discriminate between transportation modes effectively using mobile phones [9,10] or visual tracking [11]. Machine learning techniques have been used extensively to build detection models and have shown high accuracy in determining transportation modes. Supervised learning methods such as K-Nearest Neighbor (KNN) [12], Support Vector Machines (SVMs) [7, 13-17], Decision Trees [3, 4, 7, 8, 14, 18], and Random Forests (RFs) [12], have all been employed in various studies.

These studies have obtained different classifying accuracies. There are several factors that affect the accuracy of detecting transportation modes, such as the *monitoring period* (positive association), *number of modes* (negative association), *data sources, motorized classes*, and *sensor positioning* [2, 12].

However, one of the most critical factors that affects the accuracy of mode detection is the machine learning framework classifier. The framework that usually uses one layer of classification algorithm as in [4, 7] could be refered to as traditional framework; whereas the hierarchal framework uses more than one layer of classification algorithm.

An additional important consideration is the domain of the extracted features. Features are generally extracted from two different domains: (1) the time domain, features of which have been used widely in many studies [4, 12, 15, 16, 19, 20]; and (2) the frequency domain, features of which have been used in some studies [7, 15]. Both methods have achieved a significant, high accuracy. Table 4. summarizes the obtained accuracies and factors for some of the aforementioned studies. Note that no direct comparison can be made between the studies listed in Table 4. because the factors considered, and the data sets used varied from study to study.

In this study, we will mainly focus in the effect of two factors on the accuracy of transportation mode detection; namely the framework classifier and extracted features.

Most of the proposed methods in the most recent studies rely on the using of the GPS data, which do not take into account the limitations of GPS information. GPS service is not available or may be lost in some areas, which results in inaccurate position information. Moreover, the GPS

system use might deplete the smartphone's battery. Thus, this paper focuses on proposing a new detection framework using machine learning techniques and extract new features based on data obtained from smartphone sensors including accelerometer, gyroscope, and rotation vector, without GPS data.

3.4 Data Set

3.4.1 Data Collection

The data set used is available at the Virginia Tech Transportation Institute (VTTI) and was collected by Jahangiri and Rakha [12] using a smartphone app (two devices were used: a Galaxy Nexus and a Nexus 4) [12]. The app was provided to 10 travelers who work at VTTI to collect data for five different modes: driving a passenger car, bicycling, taking a bus, running, and walking. The data were collected from GPS, accelerometer, gyroscope, and rotation vector sensors and stored on the devices at the application's highest possible frequency. Data collection was conducted on different workdays (Monday through Friday) and during working hours (8:00 a.m. to 6:00 p.m.). Several factors were considered to collect realistic data reflecting natural behaviors. No specific requirement was applied in terms of sensor positioning other than carrying the smartphone in different positions that they normally do, to make sure the data collection is less dependent on the sensor positioning. The data were collected on different road types with different speed limits in Blacksburg, Virginia, and some epochs may reflect traffic jam conditions occurring in real-world conditions. The collection of thirty minutes of data over the course of the study for each mode per person was considered sufficient.

For the purpose of comparing data with data from previous studies [2, 12], the extracted features were considered to have a meaningful relationship with different transportation modes. Furthermore, features that might be extracted from the absolute values of the rotation vector sensor were excluded. Additionally, in order to allow this framework to be implemented in cases where no GPS data were available, features that might be extracted from GPS data were also excluded.

3.4.2 Time Domain Features

From the time window t, time domain features were created by applying the measures in Table 3. These measures were applied using the measurements of the data array x_i^t and its derivative $\dot{x_i^t}$ for the i^{th} feature from time window t. This resulted in 165 time domain features:

out of the 18 measures presented in in Table 3., all the 18 measures were applied to accelerometer and gyroscope sensor values; 7 measures were applied to rotation vector sensor values; 16 measures were applied to the summation values from accelerometer and gyroscope sensors; 4 measures were applied to the summation values from rotation vector sensor. As a result, the total number of features reached $18(6) + 7(3) + 16(2) + 4(1) = 165$ features [12].

Table 3. Measurements of time domain features [12].

No.	Measure	No	Measure
1	$mean(x_i^t)$	10	$spectralEntropy(x_i^t)$
2	$max(x_i^t)$	11	$mean(\dot{x}_i^t)$
3	$min(x_i^t)$	12	$max(\dot{x}_i^t)$
4	$variance(x_i^t)$	13	$min(\dot{x}_i^t)$
5	$standard\ deviation(x_i^t)$	14	$variance(\dot{x}_i^t)$
6	$range(x_i^t)$	15	$standard\ deviation(\dot{x}_i^t)$
7	$Interquartile\ range(x_i^t)$	16	$range(\dot{x}_i^t)$
8	$signChange(x_i^t)$	17	$Interquartile\ range(\dot{x}_i^t)$
9	$energy(x_i^t)$	18	$signChange(\dot{x}_i^t)$

3.4.3 Frequency Domain Features

Jahangiri and Rakha [12] collected readings from the mobile sensors at a frequency of almost 25 Hz. Because the output samples of the sensors were not synchronized, the authors implemented a linear interpolation to build continuous signals from the discrete samples. Consequently, they sampled the constructed sensor signals at 100 Hz and divided the output of each sensor in each direction ($x, y,$ and z) into non-overlapping windows of 1-s width. Finally, the features used for mode recognition were extracted from each window. These features were mainly traditional statistics such as mean, minimum, and maximum. The use of these features achieved a good accuracy in mode recognition.

However, some information loss was expected because of the usage of the summary statistics. Summary statistics consist of some descriptive statistics analysis for variability, center tendency, and distribution, such as mean, range, and variance. Summary statistics occasionally fail to detect the correlations, and extract optimal information and define probabilities [21, 22].

Since each window is considered as a signal in the time domain, we transferred each signal into the frequency domain using the short-time Fourier transform. Fourier transform converts the time function into a sum of sine waves of different frequencies, each of which represents a frequency component. The spectrum of frequency components is the frequency domain

representation of the signal. Further, the component frequencies, which are spread across the frequency spectrum, are represented as peaks in the frequency domain. These peaks represent the most dominant frequencies in the signal. However, a frequency domain can also include information on the phase shift that could be applied to each sinusoid in order to be able to recombine the frequency components to recover the original time signal. In that sense, after transforming the time domain signal to the frequency domain and neglecting the phase information, we visually inspected the resultant spectrum and found that most of the information was provided by the first 20 resulted components, which means the highest 20 magnitudes of that signal in the frequency domain.

In this study, we used the magnitude of these 20 components as the new frequency independent features. Transforming the time domain into the frequency domain not only adds new transferred features from an original space (i.e., time) to a new space (i.e., frequency), but also imposes more control on the loss of information. While the time domain represents the signal changes over time, the frequency domain adds to the time domain features: how much of the signal lies within each given frequency band over a range of frequencies. As a result, some of the expected loss in the information about signal changes in the time domain features (because of the usage of the summary statistics) might be substituted by extracting features from the frequency domain. This process resulted in the addition of another 180 features extracted from the frequency domain to the data set (i.e., 345 features pooled in total).

Table 4. Summary of some past studies [2].

Accuracy (%)	Features Domain	Machine Learning Framework	Monitoring Period	No. of Modes	Data Sources	More than One Motorized Mode?	Sensor Positioning	Data Set	Study
97.31	Time	Traditional	4 s	3	Accelerometer	Yes	No requirements	Not mentioned	[16]
93.88	Frequency	Traditional	5 s, 50% overlap	6	Accelerometer	Yes/No	Participants were asked to keep their device in the pocket of their non-dominant hip	Collected from 4 participants	[15]
93.60	Time and frequency	Traditional	1 s	5	Accelerometer GPS	No	No requirements	Collected from 16 participants	[7]
93.50	Time	Traditional	30 s	6	GPS, GIS[a] maps	Yes	No requirements	Collected from 6 participants	[4]
95.10	Time	Traditional	1 s	5	Accelerometer, gyroscope, rotation vector	Yes	No requirements	Collected from 10 participants	[12]
91.60	Time	Traditional	Entire trip	11	GPS, GIS maps	Yes	No requirements	Two different data sets, one of which included 1,000 participants	[20]
96.32	Time	Hierarchical	1 s	5	Accelerometer, gyroscope, rotation vector	Yes	No requirements	Collected from 10 participants	[2]

[a] GIS: Geographic Information System

3.5 Methods

This section describes the feature selection algorithm and the machine learning classifiers used in the proposed hierarchical framework.

3.5.1 K-Nearest Neighbor (KNN)

KNN is a common algorithm in supervised learning that classifies the data points based on the K nearest points. K is a user parameter that can be determined using different techniques. The test observation (i.e., y_j^{test}) is classified by taking the majority vote of the classes of the K nearest points (i.e., y_j^{train}), as shown in Equation (1) [23].

$$y_j^{test} = \frac{1}{K} \sum_{X_j^{train} \in N_K} y_j^{train} \quad \textbf{(1)}$$

where, y_j^{test} is the class of the testing data; y_j^{train} is the class of the training data; X_j^{train} is the testing data; and K is the number of classes.

3.5.2 Classification and Regression Tree (CART)

The CART algorithm was introduced in the early 1980s by Olshen, and Stone [24]. This algorithm is a type of decision tree where each branch represents a binary variable. At each split, the CART algorithm trains the tree using a greedy algorithm. Different splits are tested, and the split with the lowest cost is chosen. After many splits, each branch will end up in a single output variable that is used to make a single prediction. The CART algorithm will stop splitting upon reaching a certain criterion. The two most common stopping criteria are setting a minimum count of the training instances assigned to each leaf and choosing a pruning level that produces the highest accuracy.

3.5.3 Support Vector Machines (SVMs)

The SVM algorithm is a supervised learning technique that is used to classify the data by maximizing the gap between classes. The SVM algorithm attempts to find the hyperplane (i.e., splitter) that gives the largest minimum distance to the training data as given in Equation (2). The SVM tries to find the weight (w) that produces the largest margin around the hyperplane (see Equation (2)), while satisfying the two constraints (see Equations (3) and (4)) [25].

$$\min_{w,b,\xi} \left(\frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \right) \quad \textbf{(2)}$$

subject to:

$$y_n \left(w^T \phi(x_n) + b \right) \geq 1 - \xi_n \, , n = 1, \dots, N \quad \textbf{(3)}$$

$$\xi_n \geq 0 \, , n = 1, \dots, N \quad \textbf{(4)}$$

where,

w	Parameters to define the decision boundary between classes
C	Penalty parameter
ξ_n	Error parameter to denote margin violation
b	Intercept associated with the hyperplanes
$\phi(x_n)$	Function to transform data from X space into some Z space
y_n	Target value for n^{th} observation

3.5.4 Random Forest (RF)

Breiman proposed RF as a new classification and regression technique in supervised learning [26]. The RF method randomly constructs a collection of decision trees in which each tree chooses a subset of features to grow, and the results are then obtained based on the majority votes from all trees. The number of decision trees and the selected features for each tree are user-defined parameters. The reason for choosing only a subset of features for each tree is to prevent the trees from being correlated. RF was applied in this study to select the best subset of features to be used in classification, as this technique offers several advantages. For example, it runs efficiently on large datasets and can handle many input features without the need to create extra dummy variables, and it ranks each feature's individual contribution in the model [26, 27].

3.6 Proposed Framework

As many features could be used to discriminate between transportation modes, we applied feature selection to choose the subset of features with the highest importance. The subset of selected features, which is used in the classifiers, depends on the classified modes. This implies that the subset of features selected to discriminate between all modes will be different from the subset of features selected to discriminate between only two modes. In this study, RF was used to select the best 100-feature subset for each classifying step. Selected features were scaled so that the feature values were normalized to be within the range of $[-1, 1]$.

Figure 2 shows the importance of features in different ranks for all the modes combined and for different pairs of modes. The least important feature is ranked 0.1, the highest is ranked 2.2, and 0 when the features are not included. Figure 2 also illustrates that the most important feature of one pair of modes may be different for other pairs and that its rank within mode pairs

may also vary. It is noteworthy that the car-run and the car-walk pairs have lower scores as compared to the other (most of the features have a score of 0.5, which is shown in dark blue). It appears that the values of some selected features for pairs containing walk and run modes are more likely to overlap. RF ranks each feature's individual contribution in the model relatively, which means the overlap would affect the score of the individual features but not the overall classification accuracy using the entire subset of features. The overlap occurs between the features in some level of dimensionality and could be separable in the higher dimensions with high accuracy.

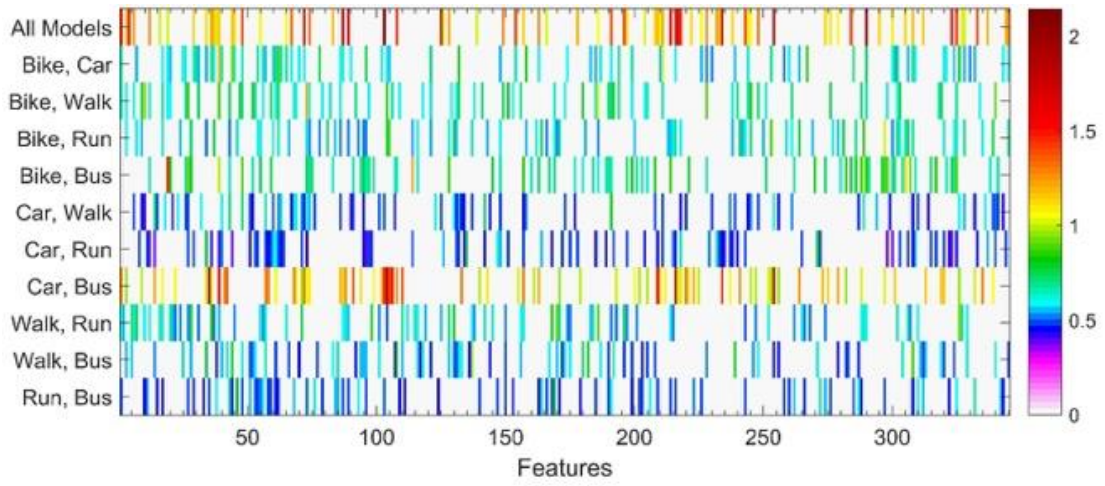

Figure 2. Importance of features for different pairs of modes.

This study proposes a new approach to detect transportation modes. Two layers are applied as a hierarchical framework. The first layer consists of only a one multiclass classifier to discriminate between the five modes, and the second layer consists of a pool of 10 binary classifiers, which are used to discriminate between only two modes. The output of each classifier (in the first or the second layer) is the probability of each mode given the test data. The first layer is trained using the RF-selected 100 features to return the corresponding modes with the highest (i) and the second highest (j) probabilities ($M^{(1)} \in \{i, j\}$). These two modes are the candidates for input to the second layer. Each classifier in the second layer is trained using a different set of RF-selected 100 features, specialized to differentiate between only the two modes of interest, to return one mode of the highest probability ($M^{(2)} \in \{i\}$). Bayesian principles are used in this framework to combine the output of the two layers. In that sense, the transportation mode with the largest posterior probability is chosen, given that the output of the first layer is the prior probability and the output of the second layer is the likelihood.

In the first layer, the probability that $M^{(1)} \in \{i\}$ is the true mode (T) in a one-layer traditional framework (i.e., $p\left(M^{(1)} \in \{i\}\middle|T\right)$) equals $P^{(1)}$. However, the probability that $M_i^{(1)}$ or $M_j^{(1)}$ (where $i \neq j$) is the true mode (T) in the two-layer proposed framework (i.e., $p\left(M^{(1)} \in \{i,j\}\middle|T\right)$) equals $P^{(1)} + \Delta$. Consequently, the proposed framework improves the potential to obtain the true mode by selecting two modes instead of only one in the first layer. The second layer consists of a pool of 10 binary classifiers (k). Thus, the probability that one mode, out of the two candidates from the first layer, is the true mode (i.e., $p(M^{(2)} \in \{i\})$) equals $c\sum_{k=1}^{10} P_k^{(2)}$. The constant c equals $\frac{1}{k}$ if the data are assumed to be balanced. Consequently, the output of the framework can be formulated using the total law of probability, as shown in Equation (5):

$$p\left(M^{(1)} \in \{i,j\}, M^{(2)} \in \{i\}, T\right) = \sum p(M^{(2)} \in \{i\}\middle|M^{(1)} \in \{i,j\}\big)\, p(M^{(1)} \in \{i,j\}|T)p(T) \qquad (5)$$

$$= \sum_{k=1}^{10} P_k^{(2)} \times (P^{(1)} + \Delta) \times c$$

$$= \left(P^{(1)} + \Delta\right)[1 - \left(1 - c\sum_{k=1}^{10} P_k^{(2)}\right)]$$

substituting $(1 - c\sum_{k=1}^{10} P_k^{(2)})$ by the summation of the errors in the second layer ($c\sum_{k=1}^{10} e_k^{(2)}$);

$$= P^{(1)} + \Delta - \left(P^{(1)} + \Delta\right)c\sum_{k=1}^{10} e_k^{(2)}$$

In fact, this term $\Delta - \left(P^{(1)} + \Delta\right)c\sum_{k=1}^{10} e_k^{(2)}$ is the difference between the output of using a one-layer traditional framework and the output of using the two-layer proposed framework. Hence, if this term is greater than zero, then there is an additional amount to probability resulting from using a one-layer traditional framework. In that case, results from the proposed framework are better than the traditional framework. This can be formulated as shown in Equation (6).

$$\Delta - \left(P^{(1)} + \Delta\right)\sum_{k=1}^{10} e_k^{(2)} > 0 \qquad (6)$$

$$\Delta > P^{(1)}(1 - c\sum_{k=1}^{10} P_k^{(2)})/c\sum_{k=1}^{10} P_k^{(2)}$$

This implies that if Δ is greater than the term $[P^{(1)}(1 - c\sum_{k=1}^{10} P_k^{(2)})/c\sum_{k=1}^{10} P_k^{(2)}]$, then the two-layer framework is beneficial. In order to examine that, we need to estimate 12 parameters from the data: $\Delta, P^{(1)}, and\ P_k^{(2)}$, where $k = 1,2, \dots ,10$. One reasonable method is to formulate the likelihood function of the classifier output as Bernoulli distribution because it is either one or zero, whereas the prior function for each the 12 parameters is formulated as a Beta distribution because it takes on any value between zero and one. This means that the prior domain is from zero to one [0,1]. Consequently, the problem can be viewed as a Beta-Bernoulli model:

$$f(y_{ij}|p_j) \sim Bernouli(p_j)$$

where $y_{ij} \in \{1,0\}$ is output i for classifier j and $p_j \in [0,1]$

$$f(p_j) \sim Beta(a, b)$$

where the values of constants a and b are chosen in which the knowledge of the prior is equal (i.e., $E[p_j] = 0.5$).

$$f\left(p_j \middle| y_{j1}, y_{j1}, \dots, y_{jN_j}\right) \propto \left\{\prod_{i=1}^{N_j} p_j^{y_{ji}} \left(1 - p_j\right)^{1-y_{ji}}\right\} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_j^{a-1}\left(1 - p_j\right)^{b\cdot}$$

the above equation can be simplified as:

$$f\left(p_j \middle| y_{j1}, y_{j1}, \dots, y_{jN_j}\right) \propto \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left\{p_j^{\sum_{i=1}^{N_j} y_i + a - 1}\left(1 - p_j\right)^{N_j - \sum_{i=1}^{N_j} y_i + b - 1}\right\}$$

by removing the constants $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ from the above equation, the kernel of the posterior is a Beta distribution with the parameters shown as:

$$f\left(p_j \middle| y_{j1}, y_{j1}, \dots, y_{jN_j}\right) \sim Beta\left(\sum_{i=1}^{N_j} y_i + a\ , N_j - \sum_{i=1}^{N_j} y_i + b\right)$$

From the above equation we can estimate the expectation $E\left[f\left(p_j \middle| y_{j1}, y_{j1}, \dots, y_{jN_j}\right)\right]$ of each parameter, as shown in Equation (7):

$$E\left[f\left(p_j \middle| y_{j1}, y_{j1}, \dots, y_{jN_j}\right)\right] = \frac{\sum_{i=1}^{N_j} y_i + a}{a + b + N_j} \qquad (7)$$

Each of the required parameters can be estimated using Equation (7). However, Equation (5), Equation (6), and the corresponding results are based on a two-layer framework. As the

number of layers in the framework increases, more parameters are required to be estimated and the model will be relatively more complicated. In addition, adding layers to the framework would increase the computational time. Yet, this does not mean that adding layers is costly, so we recommend first estimating the parameters related to the number of layers one will choose, then decide upon that.

3.7 Data Analysis and Results

This section discusses the results of the machine learning techniques used in this study, which were developed in MATLAB.

3.7.1 K-Nearest Neighbors Algorithm (KNN)

In this study, KNN was used to identify the mode from the five possible transportation modes in the first layer and the two modes in the second layer. The optimal *K* was chosen after testing different numbers of *K* versus the overall classification accuracy. To select the best model at each value of *K*, a 10-fold cross-validation was performed, and the average highest accuracy among the 10 folds was chosen. As shown in Figure 3, a higher classification accuracy was achieved using the pooled features in the proposed hierarchical framework than using only the time domain features in the same framework. Additionally, using only the time domain features, the proposed hierarchical framework outperformed traditional KNN classification using pooled features. The optimal K was found to be 7, with a highest accuracy of 95.49%.

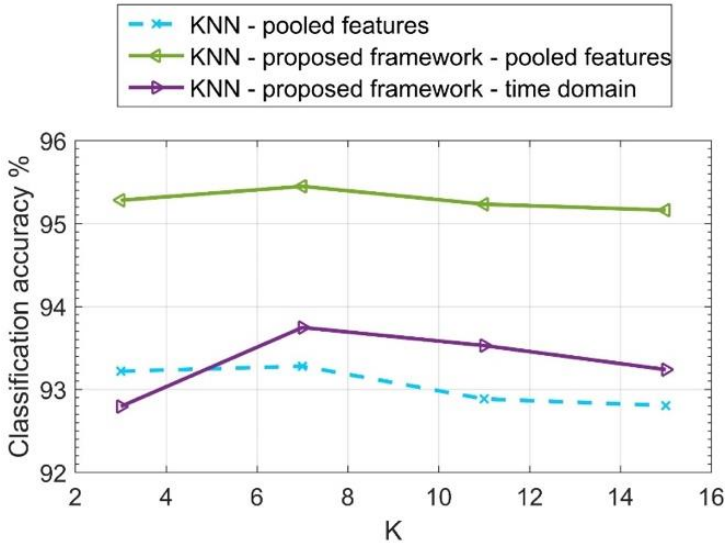

Figure 3. Classification accuracy for KNN in different cases at different neighbors.

3.7.2 Classification and Regression Tree (CART)

Ten folds for the cross-validation process were applied for each pruning level, ranging from two to 20, and the average was taken as a comparison value with other pruning levels. Figure 4 provides a comparison between time domain features, frequency domain features, and pooled features for traditional CART and CART using the proposed framework under different pruning levels. The figure shows that the proposed framework using pooled features (compared to the same applied approach using traditional CART) produces the highest accuracy among all other cases (93.52%) at six pruning levels. Figure 4 also shows that the classification accuracy of using only frequency domain features in the proposed framework approach (compared to the same approach using traditional CART) is lower than using the proposed approach and only time domain features.

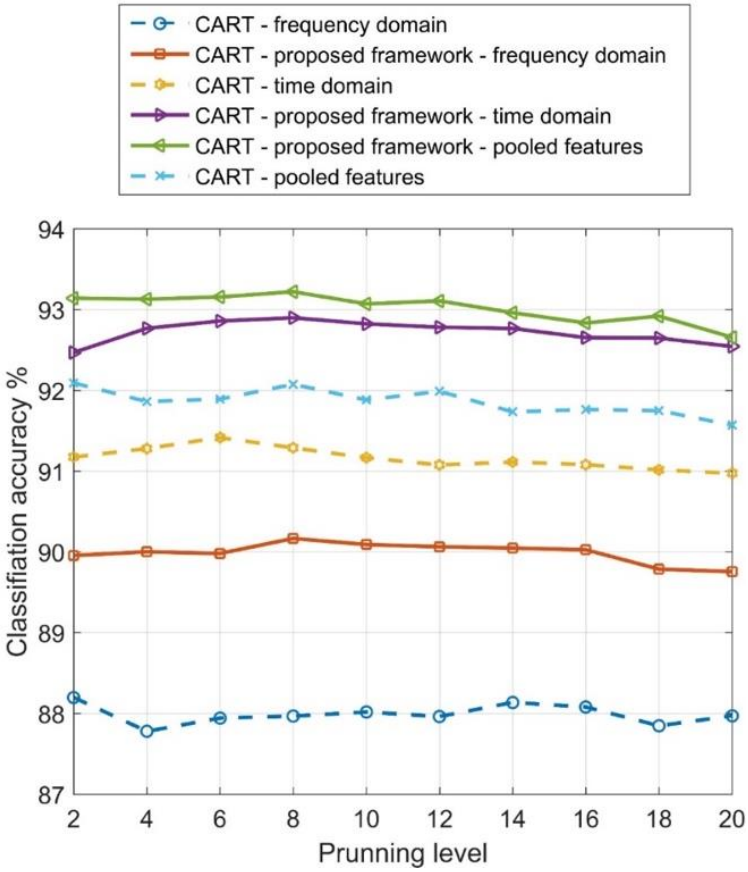

Figure 4. Classification accuracy for CART in different cases at different pruning levels.

3.7.3 Support Vector Machine (SVM)

SVM was applied in the proposed framework using time domain, frequency domain, and pooled features. A 10-fold cross-validation was applied to develop a single model. The results

show that using pooled features improved the average overall classification accuracy from 96.10% to 97.00%. The overall accuracy for using only the frequency domain features was the lowest at 93.92%. Table 5. presents the overall classification accuracy for the 10-fold testing applying the proposed SVM framework

Table 5. Overall classification accuracy for the SVM using time domain, frequency domain, and pooled features.

Fold	Time domain features (%)	Frequency domain features (%)	Pooled features (%)
1	96.04	93.78	97.12
2	96.32	93.65	97.31
3	95.88	92.90	96.88
4	96.10	93.04	97.32
5	95.98	94.01	96.76
6	96.02	94.79	96.81
7	96.38	93.62	96.98
8	95.71	94.57	96.93
9	96.32	94.52	97.01
10	96.25	94.28	96.91
Average	96.10	93.92	97.00

The confusion matrix applying SVM in the proposed framework using pooled features is given in Table 6. The precision for run mode was the highest, and the precision for bus mode was the lowest. However, the recall was the lowest for run mode and highest for bike mode.

Table 6. Confusion matrix for SVM using pooled features

		Actual					
		Bike	**Car**	**Walk**	**Run**	**Bus**	**Precision**
Predicted	**Bike**	97.13	0.52	1.17	0.40	0.58	97.33
	Car	0.66	93.57	0.16	0.13	3.06	95.88
	Walk	0.92	0.08	93.59	0.92	0.29	97.68
	Run	0.37	0.05	0.93	92.82	0.20	98.36
	Bus	0.92	2.42	0.40	0.32	93.11	95.81
	Recall	97.13	93.57	93.59	92.82	93.11	

3.7.4 Random Forest (RF)

We ran the RF with different numbers of trees to investigate the impact of the number of trees on the classification accuracy. A number of trees ranging from 200 to 400 was chosen, as the highest benefit was expected to be gained in this range according to previous studies (see more details in Elhenawy, Jahangiri, and Rakha; Jahangiri and Rakha [2, 12]). Applying RF in the proposed framework using pooled features resulted in the highest classification accuracy of

96.24% at 200 trees, as illustrated in Figure 5. Figure 5 also illustrates that applying RF using a traditional approach for classification with pooled features produces higher accuracy than the RF using the proposed classification framework with only time domain features.

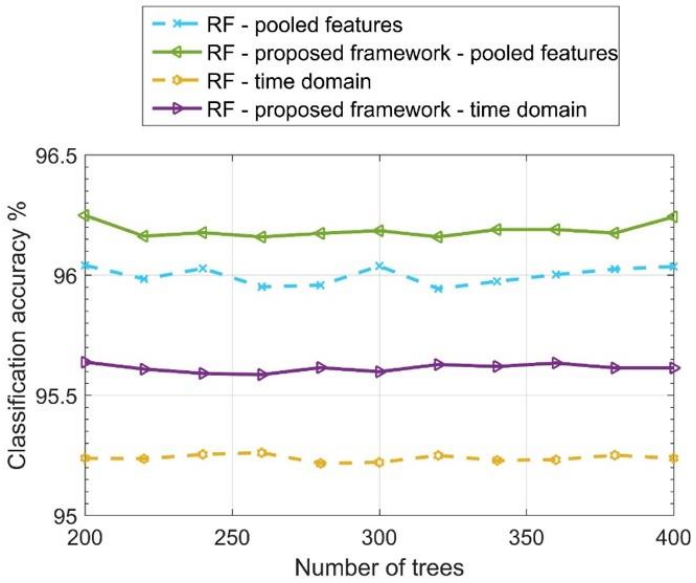

Figure 5. Classification accuracy for RF in different cases at different number of trees.

A comparison between time domain, frequency domain, and pooled features was carried out using the RF method in the proposed framework, as shown in Table 7. The results demonstrate that using the pooled features improved the overall classification accuracy from 95.61% to 96.24%.

Table 7. Overall classification accuracy for RF using time domain, frequency domain, and pooled features

Fold	Time domain features (%)	Frequency domain features (%)	Pooled features (%)
1	95.95	94.15	96.35
2	95.73	94.34	96.59
3	95.61	93.91	96.07
4	95.37	94.02	96.22
5	95.51	93.85	96.24
6	95.56	94.09	96.30
7	95.67	93.82	96.23
8	95.78	93.64	96.13
9	95.49	94.18	96.39
10	95.47	93.78	95.88
Average	95.61	93.98	96.24

Table 8. shows the confusion matrix for the RF proposed framework using pooled features. The run mode has the highest precision and the bus mode has the lowest precision.

Table 8. Confusion matrix for RF using pooled features

		Actual					
		Bike	**Car**	**Walk**	**Run**	**Bus**	**Precision**
Predicted	**Bike**	94.63	0.40	2.59	0.05	0.94	95.96
	Car	0.97	92.54	0.13	0.00	2.78	95.96
	Walk	1.87	0.10	91.74	0.25	0.70	96.92
	Run	0.75	0.05	1.47	90.39	0.57	96.96
	Bus	1.78	2.43	0.13	0.00	91.67	95.48
	Recall	94.63	92.54	91.74	90.39	91.67	

3.7.5 Heterogeneous Framework RF-SVM

We performed a heterogeneous framework in which the RF classifier was used in the first layer to classify all modes and a binary SVM classifier was applied in the second layer. The overall classification accuracy improved when using pooled features (from 96.32% to 97.02%) compared to when using only time domain features, as presented in Table 9.

Table 9. Overall classification accuracy for RF-SVM using time domain, frequency domain, and pooled features.

Fold	Time domain features (%)	Frequency domain features (%)	Pooled features (%)
1	96.51	94.26	96.96
2	96.38	94.74	96.91
3	96.52	94.78	96.86
4	96.26	94.83	96.83
5	96.44	93.71	96.97
6	96.10	95.17	96.66
7	96.12	95.30	97.36
8	96.16	94.86	97.11
9	96.33	94.49	97.39
10	96.36	94.86	97.16
Average	96.32	94.70	97.02

Table 10. and Table 11. provide the confusion matrix for applying RF-SVM in the proposed framework using time domain features and the pooled features, respectively.

Table 10. Confusion matrix for RF-SVM using time domain features

		Actual					
		Bike	**Car**	**Walk**	**Run**	**Bus**	**Precision**
Predicted	**Bike**	97.83	0.75	1.32	0.72	2.02	95.39
	Car	0.44	94.74	0.15	0.05	3.84	95.51
	Walk	1.03	0.10	97.61	0.98	0.15	97.80
	Run	0.00	0.00	0.20	97.63	0.05	99.74
	Bus	0.69	4.41	0.73	0.62	93.93	93.50
	Recall	97.83	94.74	97.61	97.63	93.93	

Table 11. Confusion matrix for RF-SVM using pooled features

		Actual					
		Bike	**Car**	**Walk**	**Run**	**Bus**	**Precision**
Predicted	**Bike**	96.12	0.34	1.17	0.06	0.71	97.79
	Car	0.69	96.81	0.16	0.01	2.85	96.27
	Walk	1.22	0.10	97.27	0.36	0.44	97.82
	Run	0.65	0.05	1.18	99.54	0.48	97.55
	Bus	1.32	2.70	0.22	0.04	95.52	95.67
	Recall	96.12	96.81	97.27	99.54	95.52	

3.8 Conclusions and Recommendations for Future Work

This study proposes a two-layer hierarchical framework classifier to distinguish between five transportation modes using new extracted frequency domain features pooled with traditionally used time domain features. The first layer contains a multiclass classifier that discriminates between five transportation modes and identifies the two most probable modes. The second layer consists of binary classifiers that differentiate between the two modes identified in the first layer. The outputs of the two layers are combined using Bayes' rule to choose the transportation mode with the largest posterior probability.

We also investigated the possibility of improving the classification accuracy using pooled features in the proposed framework by applying a number of different classification techniques, including KNN, CART, SVM, RF, and RF-SVM. The results showed that using pooled features in the proposed framework increased the classification accuracy for all of the applied classifiers. For the same data, the highest reported accuracy was 95.10% using the traditional approach for detection, whereas the proposed approach in this study achieved an accuracy of 97.02%. This implies that (a) pooling new features to be selected as classifying features increases the classification accuracy regardless of the applied approach and algorithm, and (b) applying the

proposed hierarchal framework further increases the classification accuracy. In summary, the proposed hierarchical framework outperformed the traditional approach of applying only a single layer of classifiers.

Although using pooled features increases the classification accuracy, using the new extracted features alone (i.e., frequency domain) results in a lower accuracy than only using time domain features. Transferring time domain into a new space (i.e., frequency domain) and using the magnitude of the first 20 components enhances the control on the information loss. This means that combining different features together in a big pool and then choosing the best subset of features returns better results than using one domain of features alone. The heterogeneous classifier, using RF in the first layer and SVM in the second layer, was found to produce the best overall performance.

As a future recommendation, deep analysis, such as Canonical Correlation Analysis, should be used to correlate between the features in order to obtain better coordinated results. Furthermore, future work should investigate the sensitivity of the results to the monitoring period and the potential use of GPS data.

References

[1] M. Susi, V. Renaudin, and G. Lachapelle, "Motion Mode Recognition and Step Detection Algorithms for Mobile Phone Users," *Sensors*, vol. 13, pp. 1539-62, 2013.

[2] M. Elhenawy, A. Jahangiri, and H. A. Rakha, "Smartphone Transportation Mode Recognition using a Hierarchical Machine Learning Classifier," presented at the 23rd ITS World Congress, Melbourne, Australia, 2016.

[3] X. Yu, D. Low, T. Bandara, P. Pathak, L. Hock Beng, D. Goyal, *et al.*, "Transportation activity analysis using smartphones," in *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, 2012, pp. 60-61.

[4] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu, "Transportation mode detection using mobile phones and GIS information," in *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS 2011, November 1, 2011 - November 4, 2011*, Chicago, IL, United States, 2011, pp. 54-63.

[5] A. Jahangiri, H. A. Rakha, and T. A. Dingus, "Developing a system architecture for cyclist violation prediction models incorporating naturalistic cycling data," *Procedia Manufacturing*, vol. 3, pp. 5543-5550, 2015.

[6] A. Jahangiri, H. A. Rakha, and T. A. Dingus, "Adopting Machine Learning Methods to Predict Red-light Running Violations," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, 2015, pp. 650-655.

[7] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using Mobile Phones to Determine Transportation Modes," *Acm Transactions on Sensor Networks*, vol. 6, Feb 2010.

[8] V. Manzoni, D. Maniloff, K. Kloeckl, and C. Ratti, "Transportation mode identification and real-time CO2 emission estimation using smartphones," Technical report, Massachusetts Institute of Technology, Cambridge, 2010.
[9] M. Susi, V. Renaudin, and G. Lachapelle, "Motion mode recognition and step detection algorithms for mobile phone users," *Sensors*, vol. 13, pp. 1539-1562, 2013.
[10] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, pp. 74-82, 2011.
[11] L. Wang, L. Zhang, and Z. Yi, "Trajectory predictor by using recurrent neural networks in visual tracking," *IEEE transactions on cybernetics*, vol. 47, pp. 3172-3183, 2017.
[12] A. Jahangiri and H. A. Rakha, "Applying machine learning techniques to transportation mode recognition using mobile phone sensor data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 2406-2417, 2015.
[13] L. Zhang, M. Qiang, and G. Yang, "Mobility transportation mode detection based on trajectory segment," *Journal of Computational Information Systems*, vol. 9, pp. 3279-3286, 2013.
[14] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the web," presented at the Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 2008.
[15] B. Nham, K. Siangliulue, and S. Yeung, "Predicting mode of transport from iphone accelerometer data," Tech. report, Stanford Univ, 2008.
[16] T. Nick, E. Coersmeier, J. Geldmacher, and J. Goetze, "Classifying means of transportation using mobile sensor data," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, 2010, pp. 1-6.
[17] A. Bolbol, T. Cheng, I. Tsapakis, and J. Haworth, "Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification," *Computers, Environment and Urban Systems*, 2012.
[18] P. Widhalm, P. Nitsche, and N. Brandie, "Transport mode detection with realistic Smartphone sensor data," in *2012 21st International Conference on Pattern Recognition (ICPR 2012), 11-15 Nov. 2012*, Piscataway, NJ, USA, 2012, pp. 573-6.
[19] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, p. 13, 2010.
[20] F. Biljecki, H. Ledoux, and P. Van Oosterom, "Transportation mode-based segmentation and classification of movement trajectories," *International Journal of Geographical Information Science*, vol. 27, pp. 385-407, 2013.
[21] P. G. Fedor-Freybergh and M. Mikulecký, "From the descriptive towards inferential statistics. Hundred years since conception of the Student's t-distribution," *Neuroendocrinol Lett*, vol. 26, pp. 167-171, 2005.
[22] P.-N. Tan, *Introduction to Data Mining*, India: Pearson Education, 2006.
[23] J. H. Friedman, F. Baskett, and L. J. Shustek, "A relatively efficient algorithm for finding nearest neighbors," *IEEE Trans. Comput.*, vol. 24, pp. 1000-1006, 1974.
[24] L. Olshen and C. J. Stone, "Classification and regression trees," *Wadsworth International Group*, vol. 93, p. 101, 1984.
[25] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, pp. 415-425, 2002.
[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

[27] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 14-23, 2011.

Chapter 4: Quantifying the Effect of Various Features on BSS

This chapter based on

Huthaifa I. Ashqar, Mohammed Elhenawy, Hesham A. Rakha. " Modeling Bike Counts in a Bike-Sharing System Considering the Effect of Weather Conditions." *In review*, (2016)

4.1 Abstract

The paper develops a method that quantifies the effect of weather conditions on the prediction of bike station counts in the San Francisco Bay Area Bike Share System. The Random Forest technique was used to rank the predictors that were then used to develop a regression model using a guided forward step-wise regression approach. The Bayesian Information Criterion was used in the development and comparison of the various prediction models. We demonstrated that the proposed approach is promising to quantify the effect of various features on a large BSS and on each station in cases of large networks with big data. The results show that the time-of-the-day, temperature, and humidity level (which has not been studied before) are significant count predictors. It also shows that as weather variables are geographic location dependent and thus should be quantified before using them in modeling. Further, findings show that the number of available bikes at station i at time $t-1$ and time-of-the-day were the most significant variables in estimating the bike counts at station i.

4.2 Introduction

A growing population, with more people living in cities, has led to increased pollution, noise, congestion, and greenhouse gas emissions. One possible approach to mitigating these problems is encouraging the use of bikes through bike sharing systems (BSSs). BSSs are an important part of urban mobility in many cities and are sustainable and environmentally-friendly systems. As urban density and its related problems increase, it is likely that more BSSs will exist in the future. The relatively low capital and operational cost, ease of installation, existence of pedal assistance for people who are physically unable to pedal for long distances or on difficult terrain, and better tracking of bikes are some of the properties that strengthen this prediction [1].

One of the first BSSs in the United States came into existence in 1994 with a small bike sharing program in Portland, which had only 60 bicycles available for public use. At present, although the BSS experience is still relatively limited, many cities, such as San Francisco and New York, have launched programs to serve users using different payment structures and conditions. One of the largest information technology (IT)-based systems, based in Montreal, Canada, is BIXI (BIcycle-TaXI) that uses a bicycle as a taxi. In fact, this system, with its use of advanced technologies for implementation and management, demonstrates a shift into the fourth generation of BSSs [2].

In 2013, San Francisco launched the Bay Area BSS, a membership-based system providing 24 hours a day, 7 days a week self-service access to short-term rental bicycles. Members can check out a bicycle from a network of automated stations, ride to the station nearest their destination, and leave the bicycle safely locked for someone else to use [3]. The Bay Area BSS is designed for short, quick trips, and as a result, additional fees apply to trips longer than 30 minutes. In this system, 70 bike stations connect users to transit, businesses and other destinations in four different major areas: downtown San Francisco, Palo Alto, Mountain View, and downtown San Jose [3]. The Bay Area BSS is available to everyone 18 years and older with a credit or debit card. The system is designed to be used by commuters and tourists alike, whether they are trying to get across town during the rush hour, traveling to and from the Bay Area Rapid Transit (BART) system and Caltrain stations, or for any other daily activities [3].

This paper proposes an approach to constructing a bike count model for the San Francisco Bay Area BSS. The bike counts at each station, each of which has a finite number of docks, fluctuates. Thus, a rebalancing (or redistribution) operation must be performed periodically to meet this fluctuation. Coordinating such a large operation is complicated, time consuming, polluting and expensive [1]. Firstly, this paper attempts to quantify the effect of several variables on the mean of bike counts for the Bay Area BSS network, including the month-of-the-year, the day-of-the-week, time-of-the-day, and various weather conditions. Secondly, using the same proposed method, the paper constructs a predictive model for the bike counts at each station over the time as it is one of the key tasks to making the rebalancing operation more efficient.

In terms of the paper layout, following the introduction, this paper is organized into six sections. First, related work, focused on the proposed model in previous studies, is discussed. Next, a background of count model regression, Random Forest, and Bayesian Information Criterion are

presented. Third, the different data sets used in this study are described. In the fourth section, the details of the data analysis used to quantify the effect of various features in BSS are discussed. Next, results of constructing a predictive bike count model are provided. Finally, the paper concludes with a summary of new insights and recommendations for future bike count model research.

4.3 Related Work

The modeling of BSS data using various features, including time, weather, built-environment, transportation infrastructure, etc., is an area of significant research interest. In general, the main goals of data modeling are to boost the redistribution operation [4-6], to gain new insights into and correlations between bike demand and other factors [7-10], and to support policy makers and managers in making good decisions [7, 11]. Generally, the main approach to modeling and predicting bike sharing data is regression count modeling. A recent paper modeled the demand for bikes and return boxes using data from the BSS Citybike Wien in Vienna, Austria. The influence of weather (temperature and precipitation) and full/empty neighboring stations on demand was studied using different count models (Poisson, Negative Binomial [NB] and Hurdle). The authors found that although the Hurdle model worked best in modeling the demand of bike sharing stations, these models were complex and might not be ideal for optimization procedures. They also found that NB models outperformed Poisson models because of the dispersion in the data (to be discussed later) [9]. However, an early study used count series to predict the stations' usage based on Poisson mixtures, providing insight into the relationship between station neighborhood type and mobility patterns [12].

In a study by Wang et al., log-linear and NB regression models were used to estimate total station activity counts. The factors used included: economical, built-environment, transportation infrastructural and social aspects, such as neighborhood sociodemographic (i.e., age and race), proximity to the central business district, proximity to water, accessibility to trails, distance to other bike share stations, and measures of economic activity. All the variables were found to be significant. The Log-likelihood was used as a measure of the goodness of fit of the Poisson and NB models [8]. Linear least squared regression with data from the on-the-ground Capital BSS was implemented in another paper to predict station demand based on demographic, socioeconomic, and built-environment characteristics [7].

Several studies used methods other than count models to model BSS data. A multivariate linear regression analysis was used in another study to study station-level BSS ridership. That study investigated the correlation between BSS ridership and the following factors: population density; retail job density; bike, walk, and transit commuters; median income; education; presence of bikeways; nonwhite population (negative association); days of precipitation (negative association); and proximity to a network of other BSS stations. The authors found that the demographic, built environment, and access to a comprehensive network of stations were critical factors in supporting ridership [10].

A study by Gallop et al. used continuous and year-round hourly bicycle counts and weather data to model bicycle traffic in Vancouver, Canada. The study used seasonal autoregressive integrated moving average analysis to account for the complex serial correlation patterns in the error terms and tested the model against actual bicycle traffic counts. The study demonstrated that the weather had a significant and important impact on bike usage. The authors found that the weather data (namely temperature, rain, humidity, and clearness) were generally significant; temperature and rain, specifically, had an important effect [13].

It is also worth noting that some studies used methods other than regression to either model BSS data or to develop new insights and understandings of BSSs (see [4, 11]). For example, a mathematical formulation for the dynamic public bike-sharing balancing problem was introduced using two different models: the arc-flow formulation and the Dantzig-Wolfe decomposition formulation. The demand was computed by considering the station either a pickup or delivery point, with a real-time and length period between two stations [4].

4.4 Methods

4.4.1 Count Models

In the model used for this study, the outcomes y_i (bike count in our prediction model) are discrete non-negative integers, and they represent the number of available bikes at a specified time at each station in the network. Count models based on generalized linear models (GLMs) were applied. Specifically, two models were used to predict the bike count in the network: the Poisson regression model (PRM) and the Negative Binomial regression model (NBRM). Following are brief descriptions of these two models; more details can be found in the literature [14, 15].

4.4.2 Poisson Regression Model (PRM)

In the PRM, each observation i is allowed to have a different mean μ, where μ_i is estimated from recorded characteristics. The PRM assumes that y has a Poisson distribution, and its logarithm (i.e., link function) can be modeled as a linear combination of parameters. However, the Poisson distribution assumes that the mean and variance are equal $Var(y) = \mu$. If this condition is not met, there is an over-dispersion in the data, implying that more complex models need to be applied. The probability density for the PRM is

$$f(y, \mu) = \frac{\exp(-\mu)\,\mu^y}{y!} \quad \textbf{(8)}$$

The GLM of the mean μ on a predictor vector x_i is formulated as

$$\log(\mu_i) = \beta_i x_i^T \quad \textbf{(9)}$$

where β_i are the estimated regression coefficients and $\log(\mu_i)$ is the natural logarithm.

4.4.3 Negative Binomial Regression Model (NBRM)

The NBRM is considered a generalization of PRM. It is based on a Poisson-gamma mixture distribution that assumes that the count y_i is dependent on two parameters: the mean μ_i and some dispersion parameter θ. It basically loosens the assumption in PRM that the variance is equal to the mean and adjusts the variance independently. In fact, the Poisson distribution is a special case of the Negative Binomial distribution. The probability density for the NBRM is

$$f(y, \mu) = \frac{\Gamma(y + \theta)}{\Gamma(\theta) y!} \frac{\mu^y \theta^\theta}{(\mu + \theta)^{y+\theta}} \quad (10)$$

The GLM of the mean μ on the predictor vector x_i is formulated as

$$\log(\mu_i) = \beta_i x_i^T \quad (11)$$

where β_i are the estimated regression coefficients and $\log(\mu_i)$ is the natural logarithm.

4.4.4 PRM vs NBRM

The Poisson distribution assumes that the mean and variance are the same. However, occasionally, the data shows that the variance might be higher or lower than the mean. This

situation is called over-dispersion/under-dispersion and NBRM is able to accommodate these cases. The NB distribution has an additional parameter to the Poisson distribution, which adjusts the variance independently from the mean. In fact, the Poisson distribution is a special case of the negative binomial distribution. Thus, the PRM and the NBRM have the same mean structure, but the NBRM has one parameter more than the PRM to regulate the variance independently from the mean. As Cameron and Trivedi explain, "*if the assumptions of the NBRM are correct, the expected rate for a given level of the independent variables will be the same in both models. However, the standard errors in the PRM will be biased downward, resulting in spuriously large z-values and spuriously small p-values*"[14, 16].

4.4.5 Random Forest (RF)

One of the characteristics of this type of data set is that it is often very large. It is therefore crucial to implement machine learning to identify potential explanatory variables [11]. Moreover, when a model contains a large number of predictors it becomes more complex and overfitting can occur. To avoid this, the Random Forest (RF), as introduced by Breiman in 2001 [17], was applied. The RF creates an ensemble of decision trees and randomly selects a subset of features to grow each tree. While the tree is being grown, the data are divided by employing a criterion in several steps or nodes. The correlation between any two trees and the strength of each individual tree in the forest, also known as the forest error rate in classifying each tree, affect the model. Practically, the mean squared error of the responses is used for regression. The RF method randomly constructs a collection of decision trees in which each tree chooses a subset of features to grow and, then, the results are obtained based on the majority votes from all trees. The number of decision trees and the selected features for each tree are user-defined parameters. The reason for choosing only a subset of features for each tree is to prevent the trees from being correlated.

The fact that in the RF each tree is constructed using a different bootstrap sample from the original data ensures that the RF extracts an unbiased estimate of the generalization error. This is called the OOB (out-of-bag) error estimate, which can be used for model selection and validation without the need for a separate test. The OOB was used to validate the significance of the subsequent inference of each parameter in this study. The RF technique offers several advantages. For example, it offers protection against the impact of collinearity between predictors by building bagged tree ensembles and randomly choosing a subset of features for each tree in a random forest;

it runs efficiently with a large amount of data and many input variables without the need to create extra dummy variables; it can handle highly nonlinear variables and categorical interactions; and it ranks each variable's individual contributions in the model. However, RF also has a few limitations. For instance, the observations must be independent, which is assumed in our case. Moreover, model interpretation after averaging many tree models is generally more difficult than interpreting a single-tree model. However, this is not relevant to our model, as it was used only for ranking the predictors. For more details see [17-19].

In this study, RF was used as a technique to rank the effect of the different parameters in the model. This rank was used as a systematic guide in the forward step-wise technique. Performing a direct stepwise regression for a BSS is difficult, as there are many predictors involved in the process, which is time consuming, expensive, and requires expensive statistical software (for example, see [7]). Therefore, we employed the Bayesian Information Criterion (BIC) (discussed in the next section) to choose the most accurate model while maintaining model simplicity. We started by modeling the most important parameters using RF as the only explanatory variable (aka the regressor). Then, forward step-wise regression was applied, and the log-likelihood was found and applied to determine the accumulated BIC.

4.4.6 Bayesian Information Criterion (BIC)

BIC was the criterion selected to compare between models following a forward step-wise regression guided by the results of RF. In general, the model with the lowest BIC is preferred. Adding predictors may increase the log-likelihood, leading to overfitting, and log-likelihood does not take into account the number of predictors. BIC makes up for the number of predictors in the model by introducing a penalty term. Given that $\hat{L}$ is the maximum likelihood, n is number of observations, and k is the number of predictors, BIC is defined as [20]

$$BIC = -2.\ln \hat{L} + k.\ln(n) \quad \textbf{(12)}$$

As shown in Equation (5), $k.\ln(n)$ is the term to account for the number of predictors in each model.

4.5 Data Set

This study used anonymized bike trip data collected from August 2013 to August 2015 in the San Francisco Bay Area as shown in Figure 7 [21]. This study used two data sets. The first data set included the station ID, number of bikes available, number of docks available, and time of recording. The time data included the year, month, day-of-the-month, time-of-the-day, and minutes at which an incident was recorded. As an incident was documented every minute for 70 stations in San Francisco over 2 years, this data set contained a large number of recorded incidents. This data set was exposed to a change detection process to determine times when a change in bike count occurred at each station. From this data set, as a result of pre-processing, the station ID, number of bikes available, month, day-of-the-week, and time-of-the-day were extracted for use as a feature. Time-of-the-day is considered as the time resolution of the bike counts and was regressed as 0:23; i.e. hours in a day. Subsequently, each station's ZIP code was assigned and input to the set. Figure 6 shows a histogram of the bike counts for all stations resulting from the change detection process. The histogram is considerably skewed to the right, which means that the mean, median, and mode are markedly different, indicating a dispersion in the counts.

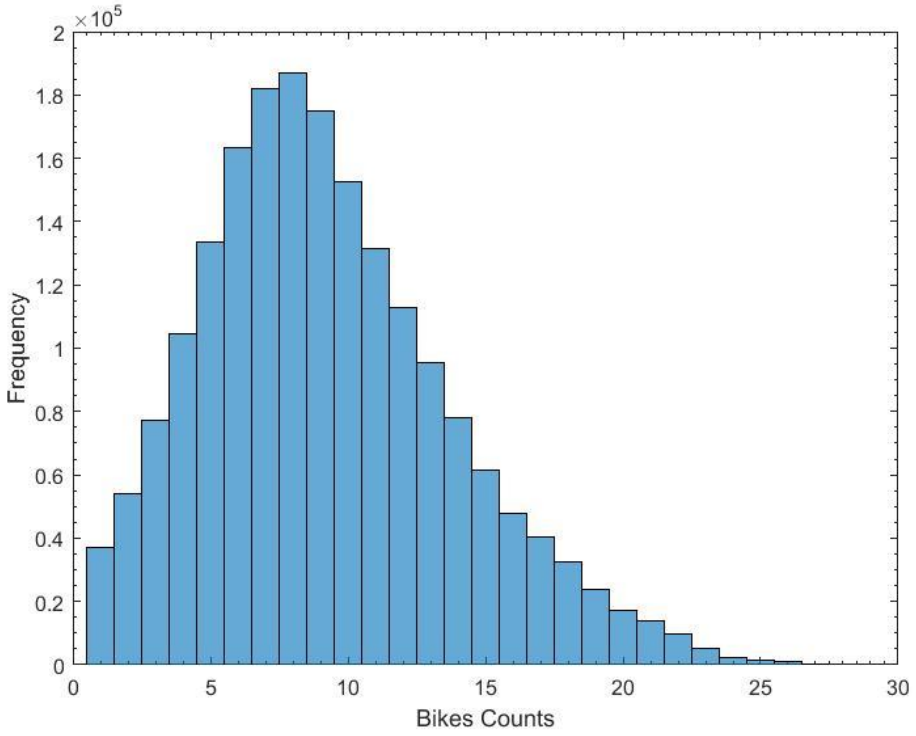

Figure 6. Histogram of the bike counts

The second data set contained different attributes: the date (in month/day/year format), ZIP code, and other variables describing the daily weather for each ZIP code over the 2-year period. Daily weather data at each ZIP code contained information about the temperature, humidity, dew level, sea level pressure, visibility, wind speed and direction, precipitation, cloud cover, and events for that day (i.e., rainy, foggy or sunny). The minimum, maximum, and mean of the first six attributes of the weather information were recorded in this data set. This data set was used to match the daily weather attributes with the first data set utilizing the two mutual attributes between them: date and ZIP code. The matched weather data was concatenated with the first data set.

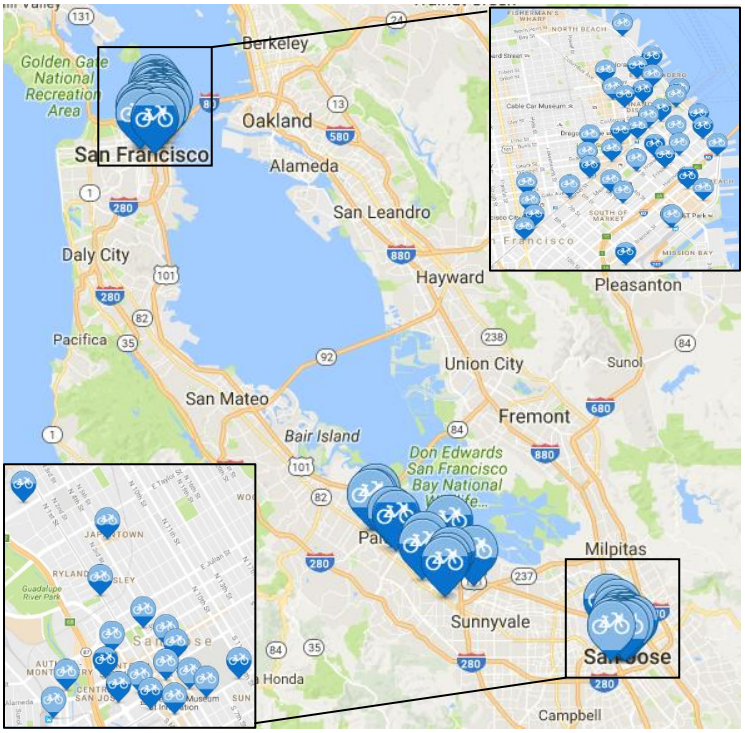

Figure 7. Stations map [3]

4.6 Data Analysis and Results

The following subsections present the methodology and the results of the data analysis. In implementing the count regression models—Poisson and Negative Binomial, RF, and BIC—MATLAB was used.

4.6.1 Problem Definition and Formulation

In quantifying the effect of various features on the system, we assumed that there is no interaction between the 70 stations and, thus station dummies were used to set up the model in this

section for two reasons: (1) the main contribution of this section is to introduce an effective and fast, but also accurate and reasonable, approach to quantifying the effect of various features on bike counts at different stations. (2) One of the important contributions of this study is to investigate the possibility of pooling all of the variables in one model instead of developing a model for each of the 70 stations. This method could be reasonable and effective in cases of large networks with big data and various variables, and also, not needing a very high estimating accuracy at specific stations. This shall depend on the task and the level of accuracy that it is needed by the operator. In practice, operators at the strategic level would use the estimation of mean bike counts with no interactions between stations to cluster the stations and/or to determine if the number of docks is sufficient in some stations. Moreover, in some cases it would be used to predict the occupancy trends of the stations to improve the quality of the service and make it more reliable for the users [22].

As we assumed there was no interaction between the 70 stations, the $\log(\mu)$ of the bike count in each station might be represented as parallel hyperplanes. In order to construct one model containing all the stations instead of a model for each station, 69 indicator variables were coded as the 70 stations in the network, which implies that Station 1 is the reference in the model intercept. Similarly, 11 indicators were coded for the 12 months with January as a reference, six indicators for the seven days of the week were coded with Sunday as a reference, and two indicators for the events in the day were coded with sunny as a reference. All of these indicators were pooled in one model. If there was no significant difference between two of the parameters (say for example β_1 and β_2), this meant that the corresponding two parallel hyperplanes (Station 1 and Station 2) were very close to each other and the predicted $\log(\mu)$ of the bike count was the same for the two stations to an acceptable level of accuracy.

The first step in understanding the bike count's behavior was to regress all the available predictors to generate a full model. To that end, the PRM and NBRM were applied. The next step was using RF to rank the predictors in the full model based on the OOB error. Forward step-wise regression was then used to fit several models that were constructed as a result of the RF. Finally, BIC was used to select the best model, or, in other words, the best subset of predictors to construct this model.

However, this subset of predictors still had to be evaluated to determine whether they were reasonable. To accomplish this, all the parameters were examined, and it was determined which

were most acceptable. Different stations, month-of-the-year, day-of-the-week, and time-of-the-day were all determined to be reasonable parameters that might affect the model. From the weather information, mean temperature, mean humidity, mean visibility, mean wind speed, precipitation, and events were selected for further investigation. These parameters were selected based on subject-matter expertise, previous related studies (see for example [9, 13]), and to avoid multi-collinearity between two or more predictors. Once again, RF and forward step-wise regression were repeated, and BIC was used to compare the built models. We chose the model with the best compromise between the minimum BIC value and the consideration of the effective parameters.

Two count models were used in this section: Poisson and negative binomial. To compare them, log-likelihood was estimated to determine goodness of fit. The likelihood of a set of parameter values is equal to the probability of the observed outcomes given those parameter values [23]. Table 12 shows the log-likelihood of Poisson and negative binomial for the full model. As negative binomial was able to accommodate the over-dispersion/under-dispersion in the data, its log-likelihood was higher than Poisson's. This meant that negative binomial was better than Poisson at describing the available bikes in the network. As a result, the NBRM was selected for use in all following steps in the analysis.

Table 12. Log-likelihood of Poisson and negative binomial models

	Poisson	NB
Log-likelihood	-5.95E+06	-5.61E+06

4.6.2 Random Forest and Bayesian Information Criterion

Both RF and BIC were applied twice in this study. RF was applied on all the available predictors, constructing 111 different models. Basically, RF was implemented to sort the predictors in descending order of their relative "importance." MATLAB's manual describes this RF measurement as "*an array containing a measure of importance for each predictor variable (feature). For any variable, the measure is the increase in prediction error if the values of that variable are permuted across the out-of-bag observations. This measure is computed for every tree, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble*"[24]. Importance was utilized as a guide in forward step-wise regression using the NBRM and computing the log-likelihood following each addition. BIC was then computed from the log-likelihood.

The BIC results of this first process are presented as the orange line shown in Figure 8. As the number of inserted predictors increased in the model, the BIC value decreased, indicating a better model. The BIC curve was used to select the most influential predictors resulting in the lowest BIC value. There was no specific rule for selecting those predictors, but rather it was a trade-off between the best and most simple model. The elbow in the curve, which corresponds to 45 predictors, was chosen to achieve the best compromise. The selected subset contained features of 31 stations, 7 months, 5 days, time-of-day, and one weather variable (wind direction degree). Based on subject-matter expertise and knowledge gained from related studies, it was determined that this subset was largely unacceptable. For example, temperature, not included in the subset, was found to be significant in previous studies of modeling bike counts in [9].

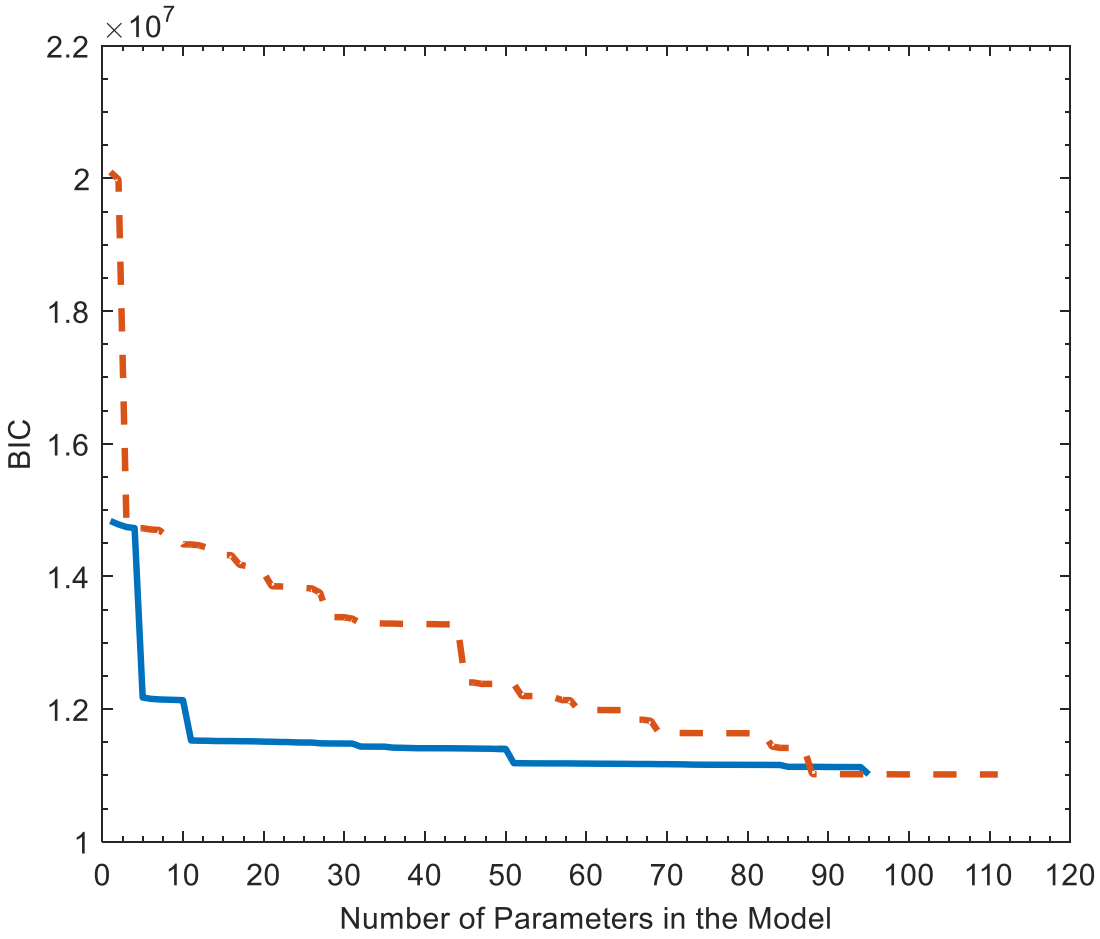

Figure 8. BIC before (orange dashed line), and after (blue solid line) feature selection process

This first conclusion led to a re-evaluation of the predictors by closely examining the weather information variables to determine any correlation among them. Again, based on expertise

and related studies, mean temperature, mean humidity, mean visibility, mean wind speed, precipitation, and events were selected as predictors. RF and BIC were again applied after the predictor selection process. The importance of the predictors resulting from the RF is shown in Figure 9 (a) and the result of the BIC following forward step-wise regression is represented by the blue line in Figure 8. As Figure 8 illustrates, selecting these features improves BIC values remarkably. This is mainly because RF obtained a different order of predictors after neglecting any features that might correlate with other parameters. For example, maximum and minimum temperatures were correlated with the mean temperature. Maximum and minimum temperatures were neglected, and the mean temperature remained.

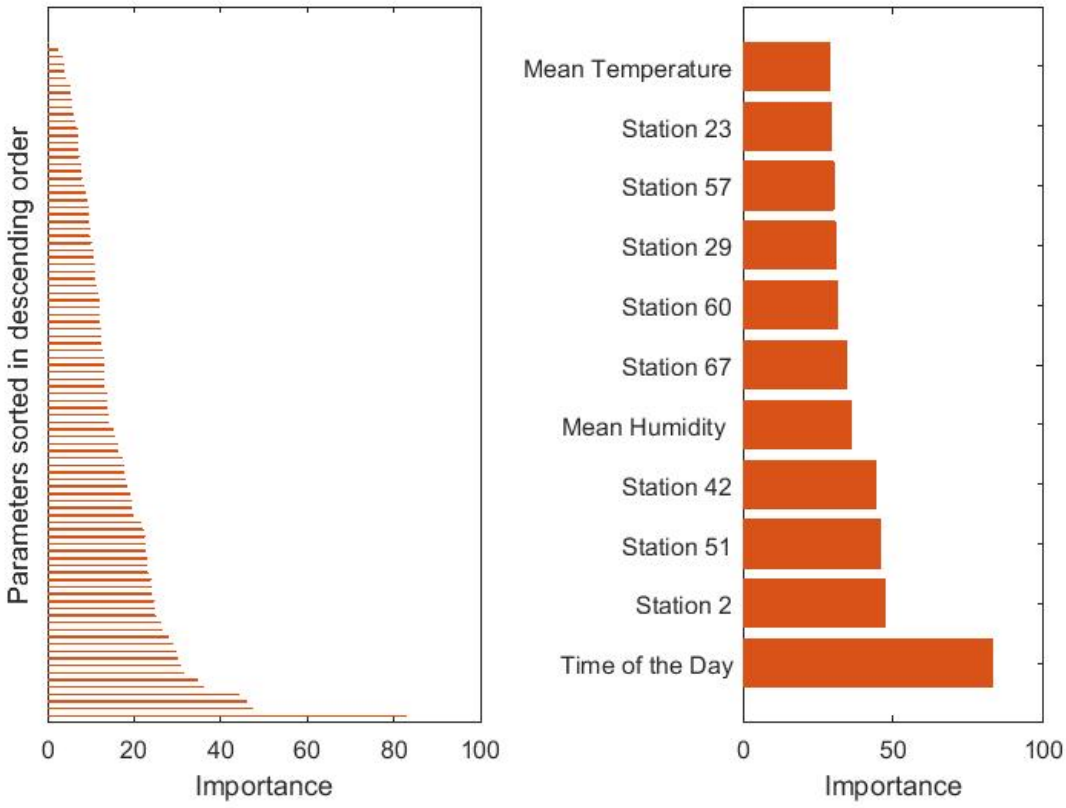

Figure 9. Importance of the predictors (a) after feature selection (b) of the first proposed solution

The BIC curve after feature selection revealed that two elbows could be selected as two proposed solutions that might achieve the best compromise: the first using 11 predictors, the second using 51 predictors. As the simplest explanation is preferable, the first solution was selected as the final model. Figure 9 (b) shows the importance of these 11 predictors, which are clearly

reasonable. Temperature and humidity turned out to be important features and have significant effects in predicting bike availability in the Bay Area Bike Share network. Bay Area is one of the most humid areas in the United States, with an average humidity of nearly 74% [25]. Humidity has been proven to be a discomfort to people, particularly during physical activities like riding a bicycle.

Although we chose the first solution, it is worth noting that if we had selected the second solution, another two weather variables (visibility and wind speed), some days of the week, and some months would be included in the 51 most important predictors. All of these predictors are also reasonable and important in predicting bike availability in the Bay Area Bike Share network. The final model is formulated as follows:

$$\log(\mu) = \beta_0 + \beta_1 ToD + \beta_2\, S2 + \beta_3 S51 + \beta_4\, S42 + \beta_5 Hu + \beta_6\, S67 + \beta_7\, S60 + \beta_8 S29 + \beta_9 S57 + \beta_{10}\, S23 + \beta_{11}\, T \quad \textbf{(13)}$$

where:

$S: Station,$

$\beta_i: Standardized\ coefficients,$

$ToD: Time - of - day,$

$Hu: Mean\ humidity\ (\%),$

$T: Mean\ temperature\ (F^o)$

This model is sufficient to estimate the mean number of bikes at each of the 70 stations producing relatively reasonable log-likelihood and BIC measures of -5.56E+06 and 1.12E+07, respectively. The log-likelihood for the reduced model is found to be higher (i.e. better) than the log-likelihood for the full model (see Table 12). When all the parameters in the full model were examined to determine which are most acceptable, we intended to exclude some parameters based on subject-matter expertise, previous related studies, and to avoid multicollinearity between the predictors, especially in the weather information. For example: mean, max, and min temperature were all regressed in the full model.

Multicollinearity in the full model is the cause of non-convergence or slow convergence of the maximum likelihood estimators, which means there is no longer a unique maximum point (i.e. peak) in the likelihood function; instead, there is a ridge [26, 27]. It appears that with collinear variables, the value of the parameter estimates fluctuates with no corresponding change in the log-

likelihood. When we avoided the multicollinearity in the reduced model, the maximum likelihood estimator converged, and increased (i.e. improved) the corresponding log-likelihood.

Although the model was set up using station dummies, it does not imply that the model could only be used to estimate the mean bike counts for the entire network. Rather, this model can be used to estimate the mean bike counts at each of the 70 stations. If one is interested in estimating the mean bike counts at Station 60, for example, then the model will be:

$$\log(\mu) = \beta_0 + \beta_1 ToD + \beta_5 Hu + \beta_7 S60 + \beta_{11} T \quad (14)$$

and all other station covariates in the model equal zero. However, if one would like to estimate the mean bike counts at Station 50 that is not included in the reduced model, then the model will be:

$$\log(\mu) = \beta_0 + \beta_1 ToD + \beta_5 Hu + \beta_{11} T \quad (15)$$

This implies two inferences, as follows: (1) There is no significant effect in including the station parameter in the model given that the mean bike counts at Station 50 is determined by three variables, namely: the time-of-the-day, the humidity level, and the ambient temperature. In other words, there is no significant difference between Station 50 and the reference station parametrized in the interception (i.e. Station 1) (2) There is no considerable difference between estimating the mean bike counts of Station 50 and, for example, Station 40 (also not included in the model), for the same time-of-the-day, humidity level, and ambient temperature. This is because the corresponding two parallel hyperplanes for Station 50 and Station 40 are very close to each other and the estimated $log(\mu)$ of the mean of bike counts is the same for the two stations to an acceptable level of accuracy.

Table 13 shows the estimated parameter values for the NB Model of mean bike counts in the studied network. It shows also that all the parameters are significant since the p-values are approximately equal to zero.

Table 13. Estimated parameter values for the NB model for bike availability in the network

	Estimate	P-value
Intercept	2.226865	< 0.0001
Time-of-the-day	-0.00050	< 0.0001
Station 2	0.467929	< 0.0001
Station 51	0.411846	< 0.0001
Station 42	0.290969	< 0.0001
Humidity	0.000516	< 0.0001
Station 67	0.428846	< 0.0001
Station 60	0.186177	< 0.0001
Station 29	2.56E-01	< 0.0001
Station 57	0.217112	< 0.0001
Station 23	0.290833	< 0.0001
Temperature	-0.0013	< 0.0001

4.6.3 Bike Count Modeling for Each Station

In this section, we use the proposed approach to modeling bike counts at each bike-sharing station using the NBRM as it appeared to outperform the PRM. Since the number of available bikes at a station, which has a finite number of docks, fluctuates, a repositioning (or redistribution) operation must be performed periodically. Coordinating such a large operation is complicated, time-consuming, polluting, and expensive [1]. Modeling the bike count at each station considering various features is one of the key tasks to making this operation more efficient. This task is the full prediction problem that would help planners make decisions such as determining the stations that need rebalancing over the entire day, and considering relocation of underused stations (or building new ones) to serve busy areas in the network. NBRM was applied to create predictive models to predict the bike counts at each of the 70 stations of the Bay Area BSS network. For each model, we used the proposed method to quantify the effect of different variables and then selected the most accurate model while maintaining model simplicity. The RF was used to rank the effect of the different parameters in the model. This rank was used as a systematic guide in the forward step-wise regression. The subset of variables includes 25 features for each station, including: month-of-the-year, day-of-the-week, time-of-the-day, and some weather variables. The weather

information contains mean temperature, mean humidity level, mean visibility level, mean wind speed, precipitation intensity, and events of fog and snow.

In [28], we studied the effect of using bike count memory data at station i as a prediction variable by ranging the memory (of 15 minutes) from $t-1$ to $t-7$, in which t is the model without including any memory data. Results showed that memory data beyond $t-1$ had a relatively small effect on bike count prediction. It seems that they do not add further explanation for the response's variability. As a result, in addition to the abovementioned subset of variables, we also added the number of available bikes at station i at time $t-1$ to estimate the bike counts at each station i in the network at time t.

Figure 10 shows the mean prediction error (MPE) for a randomly selected test sample that was not used in the estimation process for all the stations using the proposed method. As Figure 10 shows, the first 32 stations and the last two stations have relatively lower MPE than the other 36 stations. In this BSS, there are 70 bike stations that connect users in four main areas: downtown San Francisco, Palo Alto, Mountain View, and downtown San Jose. Stations that have relatively higher MPE are based in downtown San Francisco, which has a population density approximately 10 times higher than the population of the other three areas [29]. Moreover, we hypothesize that people tend to use public transportation, including BSS, in San Francisco more than the other three areas. The annual report of the TomTom Traffic Index of 2017 [30] indicates that drivers in San Francisco incur an average of 39% extra travel time while stuck in traffic anytime of the day, which is 7% more than what San Jose's drivers experience. This suggests that demographic and built environment variables are critical factors in predicting bike counts.

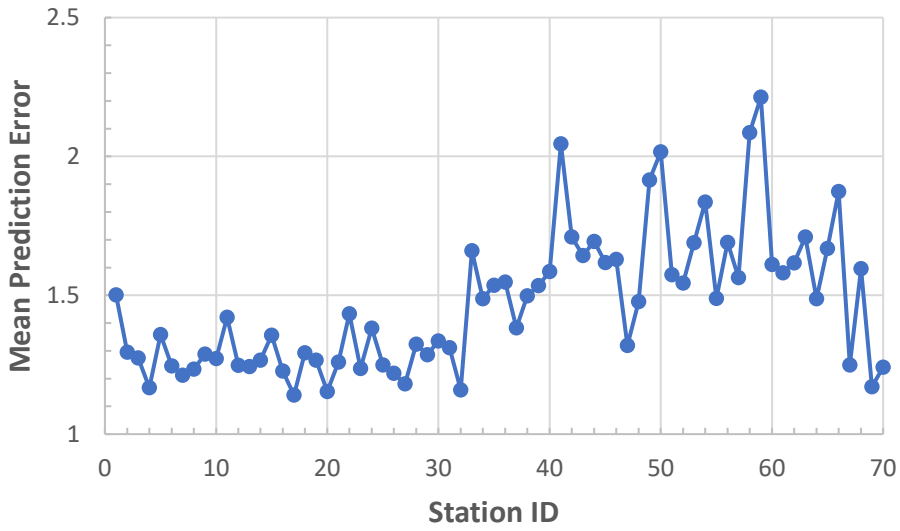

Figure 10. MPE at each station using the proposed method

Although we will present the results of modeling bike counts only at Station 3 for illustration purposes, we ran the proposed method for all the stations and the results were consistent with the presented results in terms of the variables chosen by the proposed method. The BIC curve for Station 3 in Figure 11 (a) revealed different elbows that could be selected as a proposed model. To achieve a good compromise between the BIC and the simplicity of the model, three parameters were selected to be part of the final model. The prediction results for the final model for station 3, which is shown in Figure 11 (b), is formulated as follows:

$$\log(\mu_{S3}) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 ToD + \beta_3 Hu \qquad \textbf{(16)}$$

where:

$S: Station,$

$\beta_0 = 1.24, \beta_1 = 0.1025, \beta_2 = -0.02, and\ \beta_3 = -0.005,$

$Y_{t-1}: bike\ count\ memory\ data\ at\ t-1\ (15\ minutes\ ago),$

$ToD: Time - of - day,$

$Hu: Mean\ humidity\ (\%),$

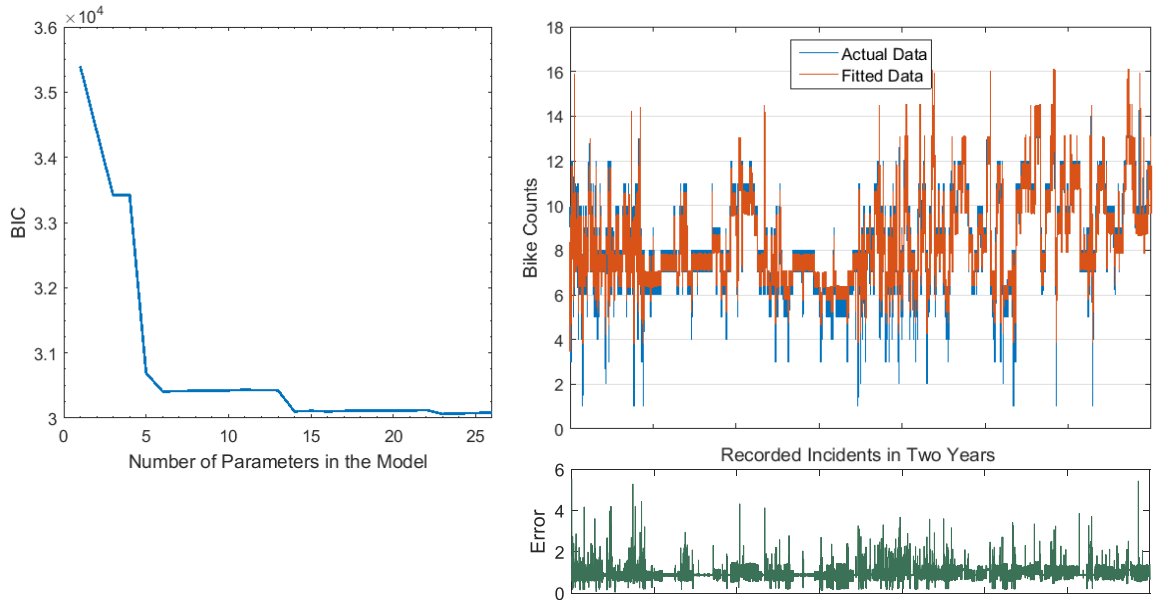

Figure 11. (a) BIC curve, and (b) fitted model for Station 3

4.7 Conclusions and Recommendations for Future Work

In this paper, we described the development of a bike availability model for the San Francisco Bay Area Bike Share program. Since the bike count estimation and prediction are still not well studied, this paper introduced an effective and fast, but also accurate and reasonable, approach to quantifying the effect of various features on bike counts at different stations. The results revealed that the bike counts changed with the month-of-the-year, day-of-the-week, time-of-the-day, and some weather variables.

NBRM and PRM were performed on the bike count data. The NBRM was ultimately chosen, as it was found to best fit the count data. However, the significance measure in NBRM (i.e., p-value) resulting from directly regressing all variables is not always an adequate measure and also depends on the order of parameters being regressed in the model, especially when there is a large number of features and if there is a possible correlation between these features. As a result, this study adopted a new method consisting of feature selection using RF. RF was run on the predictors guiding a forward step-wise regression and using the BIC to compare models. This method turned out to be an effective and reasonable approach to identify critical predictors of bike counts in the system and at each station.

The final results reveal interesting new insights. Firstly, this is the first study to use the mean humidity level as a predictor of bike counts. Results of this study demonstrate that humidity

is a significant predictor in the Bay Area Bike Share program. Further, although the precipitation level has been shown to be significant in many previous studies, the results of this study demonstrate that precipitation is not a significant predictor in the Bay Area. Over the entire year, the most common forms of precipitation in the Bay Area were light rain, moderate rain, and drizzle, none of which appeared to have a major effect on Bay Area Bike Share use. The contrast between this finding and that of previous studies indicates that particular weather information may have different significance depending on the studied geographic area.

Secondly, in investigating the effect of variables in the BSS, eight indicator variables corresponding to eight stations and one variable serving as a reference in the intercept were selected as final estimators in the model. This implies that the mean bike count data for the remaining 61 indicator variables corresponding to 61 stations are not significantly different from the mean bike count data for the reference station. The variability in bike counts of these 61 stations would not be influential if the data were employed as estimators in the regression. Nonetheless, the eight stations were different from the reference station to an extent that might largely affect the estimation if they are not considered. This is because of these station locations. For example, one station is near the main train station in Palo Alto, which is the second busiest station in the Caltrain system; another is near Yerba Buena Center for the Arts in San Francisco; one is at Union Square, which is a busy public square in the center of San Jose; and one is at the San Antonio Caltrain station in Mountain View.

Finally, the number of available bikes at station i at time $t-1$ and the time-of-the-day were found to be of the most important predictors in modeling the bike counts at each station. This means that the bike count fluctuates over the course of the day (i.e., during peak and off-peak periods). The constructed models for each station could also be used to improve the redistribution of bicycles, which is important for rebalancing the network over a period of time.

Although the adopted approach needs to be further validated by applying it to other bike count data in different geographic areas, results demonstrate that it is promising in quantifying the effect of various features in cases of large networks with big data. It is also important in the future to investigate other variables such as bikes coming from other stations and the relative location of each station.

4.8 Acknowledgements

This effort was funded by the Urban Mobility and Equitable Center and the National Science Foundation UrbComp project. The authors also acknowledge the assistance of Ahmed Ghanem and Mohammed Almannaa in data reduction and cleaning.

References

[1] P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, p. 3, 2009.

[2] S. Susan, G. Stacey, and Z. Hua, "Bikesharing in Europe, the Americas, and Asia," *Transportation Research Record*, pp. 159-167, 2010.

[3] Bay Area Bike Share. (2016). *Introducing Bay Area Bike Share, your new regional transit system*. Available: http://www.bayareabikeshare.com/faq#BikeShare101

[4] C. Contardo, C. Morency, and L.-M. Rousseau, *Balancing a dynamic public bike-sharing system*. Cirrelt, 2012.

[5] J. Schuijbroek, R. Hampshire, and W.-J. van Hoeve, "Inventory rebalancing and vehicle routing in bike sharing systems," 2013.

[6] T. Raviv, M. Tzur, and I. A. Forma, "Static repositioning in a bike-sharing system: models and solution approaches," *EURO Journal on Transportation and Logistics*, vol. 2, no. 3, pp. 187-229, 2013// 2013.

[7] D. W. Daddio, "Maximizing Bicycle Sharing: an empirical analysis of capital bikeshare usage," 2012.

[8] X. Wang, G. Lindsey, J. E. Schoner, and A. Harrison, "Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations," *Journal of Urban Planning and Development*, vol. 142, no. 1, p. 04015001, 2015.

[9] C. Rudloff and B. Lackner, "Modeling demand for bicycle sharing systems–neighboring stations as a source for demand and a reason for structural breaks," 2013.

[10] R. Rixey, "Station-Level Forecasting of Bikesharing Ridership: Station Network Effects in Three US Systems," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2387, pp. 46-55, 2013.

[11] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns," *Procedia - Social and Behavioral Sciences*, vol. 20, pp. 514-523, 2011/01/01 2011.

[12] C. Etienne and O. Latifa, "Model-Based Count Series Clustering for Bike Sharing System Usage Mining: A Case Study with the Velib' System of Paris," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1-21, 2014.

[13] C. Gallop, C. Tse, and J. Zhao, "A seasonal autoregressive model of Vancouver bicycle traffic using weather variables," *i-Manager's Journal on Civil Engineering*, vol. 1, no. 4, p. 9, 2011.

[14] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*. Cambridge university press, 2013.

[15] J. S. Long and J. Freese, *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

[16] A. C. Cameron and P. K. Trivedi, "Econometric models based on count data. Comparisons and applications of some estimators and tests," *Journal of applied econometrics*, vol. 1, no. 1, pp. 29-53, 1986.
[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
[18] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14-23, 2011.
[19] K. Matsuki, V. Kuperman, and J. A. Van Dyke, "The Random Forests statistical technique: An examination of its value for the study of reading," *Scientific Studies of Reading*, vol. 20, no. 1, pp. 20-33, 2016.
[20] E. Wit, E. v. d. Heuvel, and J. W. Romeijn, "'All models are wrong...': an introduction to model uncertainty," *Statistica Neerlandica*, vol. 66, no. 3, pp. 217-236, 2012.
[21] B. Hamner. (2016). *SF Bay Area Bike Share | Kaggle*. Available: https://www.kaggle.com/benhamner/sf-bay-area-bike-share
[22] G. M. Dias, B. Bellalta, and S. Oechsner, "Predicting occupancy trends in Barcelona's bicycle service stations using open data," 2015, pp. 439-445: IEEE.
[23] S. Johansen and K. Juselius, "Maximum likelihood estimation and inference on cointegration—with applications to the demand for money," *Oxford Bulletin of Economics and statistics*, vol. 52, no. 2, pp. 169-210, 1990.
[24] MathWorks. (2016). *Variable importance for prediction error - MATLAB*. Available: https://www.mathworks.com/help/stats/treebagger.oobpermutedvardeltaerror.html
[25] Current Results. (2016). *Most Humid Cities in USA - Current Results*. Available: https://www.currentresults.com/Weather-Extremes/US/most-humid-cities.php
[26] H. A. L. Kiers, "A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity," *Journal of Chemometrics*, vol. 12, no. 3, pp. 155-171, 1998.
[27] J. Shen and S. Gao, "A Solution to Separation and Multicollinearity in Multiple Logistic Regression," *Journal of data science : JDS*, vol. 6, no. 4, pp. 515-531, 2008.
[28] H. I. Ashqar, M. Elhenawy, and H. A. Rakha, "Network and Station-Level Bike-Sharing System Prediction: A San Francisco Bay Area Case Study," 2017.
[29] World Population Review. (2017). *California Population 2017 (Demographics, Maps, Graphs)*. Available: http://worldpopulationreview.com/states/california-population/
[30] TOMTOM. (2017). *TomTom Traffic Index 2017*. Available: http://corporate.tomtom.com/releasedetail.cfm?ReleaseID=1012517

Chapter 5: Network and Station-Level BSS Prediction

This chapter based on

Huthaifa I. Ashqar, Mohammed Elhenawy, Hesham A. Rakha. " Network and Station-Level Bike-Sharing System Prediction: A San Francisco Bay Area Case Study." *In review*, (2018)

5.1 Abstract

The paper develops models for modeling the availability of bikes in the San Francisco Bay Area Bike Share System using machine learning algorithms at two levels: network and station. Random Forest and Least-Squares Boosting were used as univariate regression algorithms, and Partial Least-Squares Regression (PLSR) was applied as a multivariate regression algorithm. The univariate models were used to model the number of available bikes at the station-level. PLSR was applied to reduce the number of required prediction models and reflect the spatial correlation between stations at the network-level. Results showed that univariate models had lower error predictions than the multivariate model. Moreover, results of the station-level analysis suggested that demographic and built environment variables were critical factors in predicting bike counts. We also demonstrated that the available bikes modeled at the station-level at time t had a notable influence on the bike count models. The multivariate model results were reasonable at the network-level, with a relatively large number of spatially correlated stations. Results also showed that station neighbors and prediction horizon times were significant predictors. The most effective prediction horizon time that produced the least prediction error was 15 minutes.

5.2 Introduction

In the next few decades, many traditional cities will be turned into smart cities, which are greener, safer, and faster. This transformation is supported by recent advances in information and communication technology (ICT) in addition to the expected fast spread of the Internet of Things (IoT) and big data analytics. Smart cities may mitigate some of the negative impacts of traditional cities, which consume 75% of the world's energy and produce 80% of greenhouse gases [1]. Smart cities have many components, including smart transportation. Smart transportation will integrate different transportation networks and allow them to work together so travelers and commuters can

enjoy seamless multi-modal trips based on their preferences. Consequently, more commuters will be inspired to use public transportation systems and many traffic-related problems such as congestion could be reduced.

A growing population, with more people living in cities, has led to increased pollution, noise, congestion, and greenhouse gas emissions. One possible approach to mitigating these problems is encouraging the use of bike-sharing systems (BSSs). BSSs are an important part of urban mobility in many cities and are sustainable and environmentally friendly. As urban density increases, it is likely that more BSSs will appear due to their relatively low capital and operational costs, ease of installation, pedal assistance for people who are physically unable to pedal for long distances or on difficult terrain, and ability to track bikes [1].

One of the first BSSs in the United States was established in 1964 in Portland, with 60 bicycles available for public use. Although BSSs are still relatively limited, at present many cities, such as San Francisco and New York, have launched BSS programs. These programs implement different payment structures, conditions, and logistical strategies. In 2013, San Francisco launched the Bay Area Bike Share System (BSS) (now called the "Ford GoBike" BSS), a membership-based system providing 24-hours-per-day, 7-days-per-week self-service access to short-term rental bicycles. Members can check out a bicycle from a network of automated stations, ride to the station nearest their destination, and leave the bicycle safely locked for someone else to use [2]. The Bay Area BSS is designed for short, quick trips, and as a result, additional fees apply for trips longer than 30 minutes. In this system, 70 bike stations connect users to transit, businesses, and other destinations in four areas: downtown San Francisco, Palo Alto, Mountain View, and downtown San Jose [2]. The Bay Area BSS is available to everyone 18 years and older with a credit or debit card. The system is designed to be used by commuters and tourists alike, whether they are trying to get across town at rush hour, traveling to and from Bay Area Rapid Transit (BART) and Caltrain stations, or pursuing daily activities [2].

This paper proposes an approach to modeling the number of available bikes at a bike share station using machine learning. Since the number of available bikes at a station, which has a finite number of docks, fluctuates, a repositioning (or redistribution) operation must be performed periodically. Coordinating such a large operation is complicated, time-consuming, polluting, and expensive [1]. Studying the number of available bikes at a station-level is one of the key tasks to making this operation more efficient. In this study, Random Forest (RF) and Least-Squares

Boosting (LSBoost) algorithms were used to build univariate prediction models for available bikes at each Bay Area bike station. However, to reduce the number of required prediction models for the entire BSS network, we also used Partial Least-Squares Regression (PLSR) as a multivariate regression algorithm.

5.3 Related work

Modeling bike sharing data is an area of significant research interest. Proposed models have relied on various features, including time, weather, the built environment, and transportation infrastructure. In general, the main goals of these models have been to boost the redistribution operation [3-5], to gain new insights into and correlations between bike demand and other factors [6-9], and to support policy makers and managers in making optimized decisions [6, 10].

Many studies have been performed at the station-level to predict the availability of bikes by using time series analysis. Rixey used a multivariate linear regression analysis to study station-level BSS ridership, investigating the correlation between BSS ridership and the following factors: population density; retail job density; bike, walk, and transit commuters; median income; education; presence of bikeways; nonwhite population (negative association); days of precipitation (negative association); and proximity to a network of other BSS stations.[9] The author found that demographics, the built environment, and access to a comprehensive network of stations were critical factors in supporting ridership.

Froehlich, Neumann, and Oliver used four predictive models to predict the number of available bikes at each station: last value, historical mean, historical trend, and Bayesian networks [11]. Two methods for time series analysis, autoregressive-moving average (ARMA) and autoregressive integrated moving average (ARIMA), have also been used to predict the number of available bikes/docks at each bike station. Kaltenbrunner et al. adopted ARMA [12]; Yoon et al. proposed a modified ARIMA model considering spatial interaction and temporal factors [13]. However, Gallop et al. used continuous and year-round hourly bicycle counts and weather data to model bicycle traffic in Vancouver, Canada [14]. That study used a seasonal autoregressive integrated moving average analysis to account for the complex serial correlation patterns in the error terms and tested the model against actual bicycle traffic counts. The results demonstrated that the weather had a significant and important impact on bike usage. The authors found that the weather data (i.e., temperature, rain, humidity, and clearness) were generally significant and that

temperature and rain, specifically, had an important effect.

However, few studies have used machine learning to model bike sharing data. One of the characteristics of transportation-related datasets is that they are often very large. It is therefore advantageous to implement machine learning to identify potential explanatory variables [10]. Moreover, when a model contains a large number of predictors, it becomes more complex and overfitting can occur. To address this, different algorithms have been used to predict bike availability in a BSS, such as random forest (RF), support vector machine (SVM), and gradient boosted tree (GBT) [15-20]. The authors of the four studies in [15, 16, 18, 19] used different machine learning algorithms to predict bike demand based on the usage record and other information about the targeting prediction time window. While the full prediction problem would be predicting bike counts at each station, the authors used machine learning to predict the bike count of the entire BSS instead. In [17], the authors used RF to classify the status of the stations only with regard to whether the station was completely full of bikes or completely empty, so users could not return a bicycle, or could not find one to rent.

This paper makes three major contributions to the literature. 1) modeling bike count prediction at the station-level using machine learning algorithms has not been studied well to date. 2) the univariate response models previously used to predict the number of available bikes at each station ignore the correlation between stations and might become hard to implement when applied to relatively large networks. This paper investigates the use of multivariate response models to predict the number of available bikes in the network. 3) Station neighbors, which are determined by a trip's adjacency matrix, are considered as significant predictors in the regression models.

5.4 Methods

5.4.1 Random Forest (RF)

Breiman proposed RF as a new classification and regression technique in supervised learning [21]. RF creates an ensemble of decision trees and randomly selects a subset of features to grow each tree. While the tree is being grown, the data are divided by employing a criterion in several steps or nodes. The forest rate error depends on the correlation between any two trees and the strength of each individual tree in the forest. Practically, the mean squared error of the responses is used for regression.

RF offers several advantages [21, 22]. For example, there are very few assumptions

attached to its theory; it is considered to be robust against overfitting; it runs efficiently and relatively quickly with a large amount of data and many input variables without the need to create extra dummy variables; it can handle highly nonlinear variables and categorical interactions; and it ranks each variable's individual contributions in the model. However, RF also has a few limitations. For instance, the observations must be independent, which is assumed in our case.

5.4.2 Least-Squares Boosting (LSBoost)

LSBoost is a gradient boosting of regression trees that produces highly robust and interpretable procedures for regression. LSBoost was proposed by Friedman as a gradient-based boosting strategy [23], using square loss $L\,(y, F) = (y - F)^2/2$, where F is the actual training and y is the current cumulative output $y_i = \beta_0 + \sum_{j=1}^{i-1} \beta_j h_j + \beta_i h_i = y_{i-1} + \beta_i h_i$. The new added training $\hat{F}$ is set to minimize the loss, in which the training error is computed as in [24]:

$$E = \sum_{t=1}^{N} \left[\beta_i h_i^t - \hat{F}^t\right] \qquad (1)$$

where $\hat{F}$ is the current residual error and the combination coefficients β_i are determined by solving $\partial E/\partial \beta_i = 0$.

In this paper, RF and LSBoost were used as univariate regression techniques to model the number of available bikes in each station at any time t. RF and LSBoost are ensemble learning algorithms, which integrate multiple decision trees to produce robust models. However, the main difference between these two algorithms is the order in which each component tree is trained. Using randomness, RF trains each tree independently, whereas LSBoost trains one tree at a time and each new added tree is set to correct errors made by previously trained trees. The ensemble model is produced by synthesizing results from the individual trees.

5.4.3 Partial Least-Squares Regression (PLSR)

PLSR was recently developed as a multivariate regression algorithm [25-29]. PLSR finds a linear regression model by projecting the predicted variables Y and the observable variables X to a new space. The basic model in the PLSR method consists of a regression between two blocks; i.e. X and Y. Furthermore, this model contains outer relations for each of the X and Y blocks, and an inner relation that links both blocks. PLSR has several advantages. For example, it is suitable when the matrix of predictors Y has more variables than observations, and when there is

multicollinearity among observable variable X values. Moreover, the PLSR method outperforms multiple linear regressions because implementing PLSR develops stable predictors. In this paper, PLSR was used as multivariate regression to reduce the number of required prediction models for the number of available bikes at any time t for the entire BSS network.

5.5 Dataset

This study used anonymized bike trip data collected from August 2013 to August 2015 in the San Francisco Bay Area, as shown in Figure 12 [30]. This study used two datasets. The first dataset includes station ID, number of bikes available, number of docks available, and time of recording. The time data include year, month, day of month, day of week, time of day, and minutes at which a record was documented. As an incident was documented every minute for 70 stations in San Francisco over 2 years, this dataset contains a large number of recorded incidents. Consequently, the size of the dataset was reduced by sampling station data once at every quarter-hour instead of once every 1 minute and obtaining the exact values without any smoothing process. This was done to reduce the complexity of the data and take a global view of bike availability in the entire network every 15 minutes.

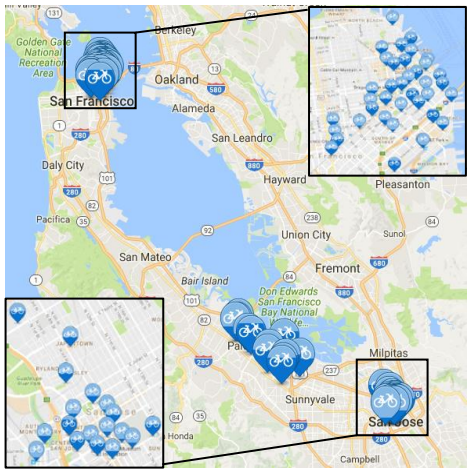

Figure 12. Stations map. (Source: Google Maps)

During the data processing phase, we found that numerous stations had recently been added to the network and others had been terminated. As a result, the dataset was cleaned by eliminating any entries missing docking station data. This reduced the number of entries from approximately 70,000 to 48,000. Each entry included the availability of bikes at the 70 stations with the associated time (month of year, day of week and hour of day) and the weather information. The weather

information included mean temperature, mean humidity, mean visibility, mean wind speed, precipitation, and events in a day (i.e., rainy, foggy, or sunny). These parameters were selected based on subject-matter expertise and previous related studies [8, 14], and were found to be significant in predicting the number of available bikes at Bay Area BSS stations [31].

Moreover, trip data included detailed information about origin station, destination station, and time of each bike trip within the BSS during the 2 years. We used the trip data to generate the BSS network adjacency matrix and found the highest 10 in-degree stations for station i, which were assigned as neighbors of station i. In other words, the neighbors of a station i were defined based on the number of trips that originated from station j, in which $j \neq i$, and ended at station i.

5.6 Data analysis and results

5.6.1 Univariate Models

RF and LSBoost algorithms were applied to create univariate models to predict the number of available bikes at each of the 70 stations of the Bay Area BSS network. The two algorithms were applied to investigate the effect of several variables on the prediction of the number of available bikes in each station i in the network, including the available bikes at station i at time t, the available bikes at its neighbors at the same time t, the month of year, day of week, time of day, and various selected weather conditions. The predictors' vector for station i at time t, denoted by X_t^i, was used in the built models to predict the log of the number of available bikes at station i at time $t + \Delta$, which is denoted by $\log(y_{t+\Delta}^i)$, where $i = 1, 2, \dots, 70$ and Δ is the prediction horizon time. The effect of different prediction horizons, Δ (range 15–120 minutes), on the performance of both algorithms was investigated by finding the Mean Absolute Error (MAE) per station (i.e., bikes/station), which can be described as the prediction error. Moreover, as the number of generated trees by RF and LSBoost is an important parameter in implementing both algorithms, we investigated its effect by changing the number of generated trees from 20 to 180 with a 40-tree step.

As shown in Figure 13 and Figure 14, the prediction errors of RF and LSBoost increased as the prediction horizon Δ increased. The lowest prediction error for both algorithms occurred at a 15-minute prediction horizon. Moreover, the prediction error of RF and LSBoost decreased as the number of trees increased until it reached a point where increasing the number of trees would

not significantly improve the prediction accuracy. Figure 13 and Figure 14 also show that a model consisting of 140 trees yields a relatively sufficient accuracy.

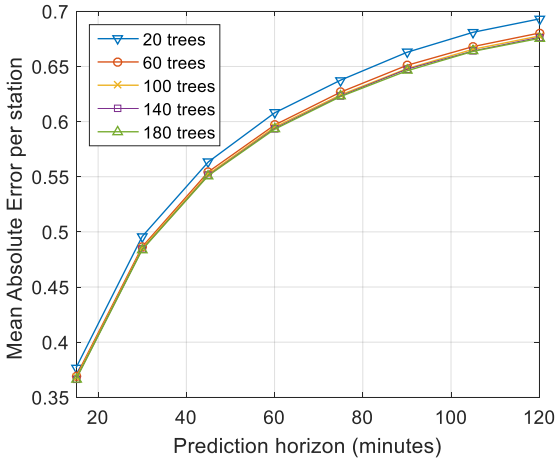

Figure 13. RF MAE at different prediction horizons and number of trees.

Comparing the two algorithms, the models produced by RF generally had a smaller prediction error than those produced by LSBoost. LSBoost is a gradient-boosting algorithm, which usually requires various regularization techniques to avoid overfitting [32]. As Figure 14 clearly shows, as the prediction horizon time increases, the prediction error increases.

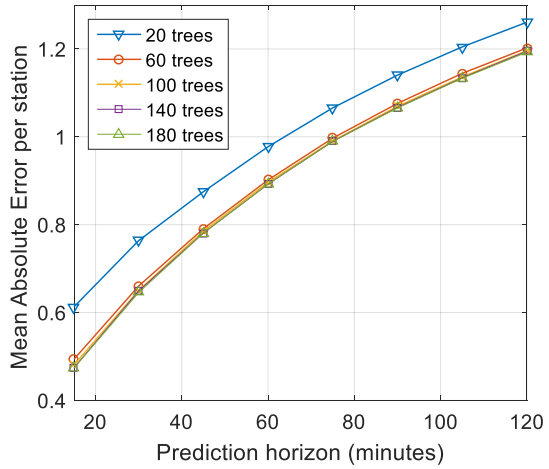

Figure 14. LSBoost MAE at different prediction horizons and number of trees.

5.6.2 Station-Level Analysis

Investigating the BSS at the station-level is the full prediction problem that would help planners make decisions such as considering the relocation of underused stations (or building new ones) to serve busy areas in the network. In order make a station-level analysis of the BSS, we

used RF, which consisted of 140 trees at a prediction horizon time of 15 minutes, to model the bike availability at each bike station in the BSS network. The behavior of the algorithm at different stations was investigated by finding the maximum absolute error ($MaxAE = max(|y^i_{t+\Delta} - \hat{y}^i_{t+\Delta}|)$ where$i = 1, 2, \dots, 70,$ and $\Delta =$ 15 minutes) at each station as shown in Figure 15. We used $MaxAE$ to explore the max prediction error that might occur in some time t at each station using 5-fold of cross validation.

As Figure 15 shows, the first 32 stations and the last two stations have relatively lower $MaxAE$ than the other 36 stations. In this BSS, there are 70 bike stations that connect users in four main areas: downtown San Francisco, Palo Alto, Mountain View, and downtown San Jose. Stations that have relatively higher $MaxAE$ are based on downtown San Francisco, which has a population density approximately 10 times higher than the population of the other three areas [33]. Moreover, hypothesize that people tends to use public transportation, including BSS, in San Francisco more than the other three areas. The annual report of TomTom Traffic Index 2017 [34] indicates that drivers in San Francisco expecting to spend an average of 39% extra travel time stuck in traffic anytime of the day, which is 7% more than San Jose's drivers. During our analysis, we visually compared the bike counts of the two groups of stations during the period of August 2013 to August 2015, finding that bike counts in stations based in downtown San Francisco were more volatile than other stations. This suggests that demographic and built environment variables are critical factors in predicting bike counts.

Moreover, some stations in that area were found to be highly unpredictable due to the high fluctuation in bike counts. Harry Bridges Plaza Station, which has the highest $MaxAE$, is an example of this type of station. Figure 16 shows the absolute error and the corresponding actual and RF-model-predicted bike counts at this station during a randomly-selected week. In fact, we found that in the future the operator company has planned a coming-soon station very near to Harry Bridges Plaza Station to increase its capacity [35]. As Figure 16 shows, the fluctuation occurs suddenly in one single observation, during which actual bike counts may rise or fall steeply in a very short period. This high fluctuation in bike counts at these stations can be divided into two types: 1) fluctuation due to the periodic redistribution operation to rebalance bike counts; 2) fluctuation due to the high incoming/outgoing demand in the station within a relatively short period. When we studied the area around Harry Bridges Plaza Station, we hypothesize that this high incoming/outgoing demand comes from it being an open air area at the end of market and

restaurants, where artists, skaters, tourists and others congregate to enjoy the happenings and beautiful scenery [36].

It was difficult to classify all the fluctuating incidents as type one or type two. Nonetheless, the first type of fluctuation can be addressed by adding *'rebalancing difference'* to the predicted bike counts ($\hat{y}^{i}_{t+\Delta}$) if the number of redistributed number and the time of redistribution operation are available. However, this solution is not applicable in all cases.

To further address fluctuation, we then studied the effect of using bike count memory data at station $\boldsymbol{i}$ as a prediction variable by ranging the memory from $\boldsymbol{t-1}$ ***to*** $\boldsymbol{t-7}$, in which $\boldsymbol{t}$ in Figure 17 is the model without including any memory data. Results in Figure 17 show that memory as a prediction variable had a relatively small effect on bike count prediction and could not be used to relax fluctuation in this case study.

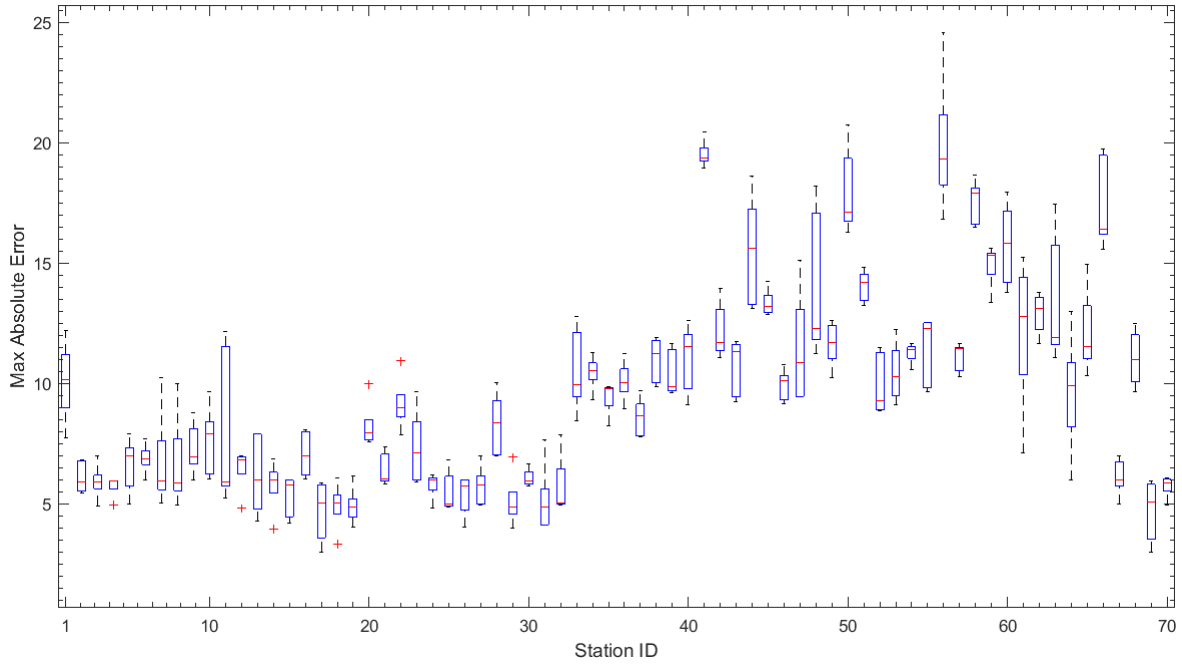

Figure 15. MaxAE at each station using RF.

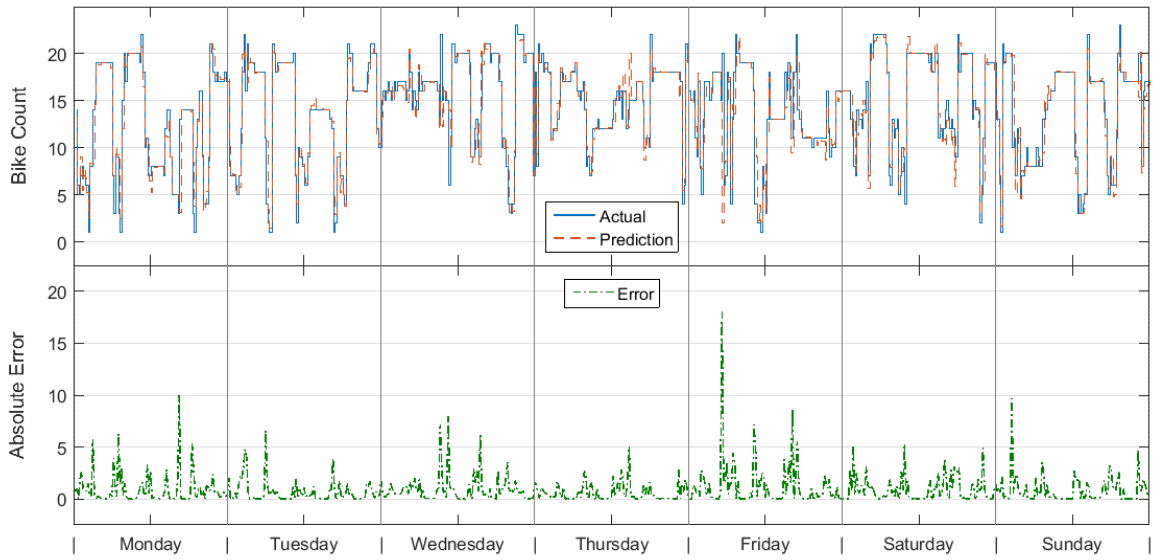

Figure 16. AE and bike availability at Harry Bridges Plaza Station during a selected week

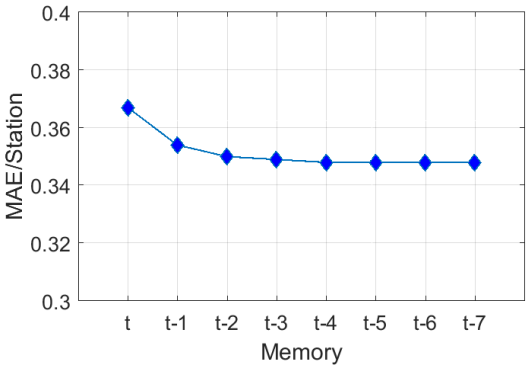

Figure 17. MAE using RF at different prediction horizons.

This finding suggested that it is crucial to use the available bikes at station-level at time t, and, to a smaller extent, the available bikes at its neighboring stations at the same time t, as variables to predict the number of available bikes in each station i. Although the predictive model still lags by one step, using these variables has a notable influence that limits fluctuation by forcing the predicted bike count model to follow the actual bike count instantaneously when fluctuation occurs in the 15-minute prediction horizon time. This process is shown in Figure 16.

5.6.3 Multivariate Model

PLSR was used as a multivariate regression to reduce the number of required prediction models for bike stations in the BSS network. When a BSS network has a relatively large number of stations, tracking all the models for each bike station becomes complex and time-consuming.

For that reason, we examined the adjacency matrix of the Bay Area BSS network and found that the network can be divided into five regions, as shown in Figure 18. The regions that resulted from the adjacency matrix were found to consist of bike stations that shared the same ZIP code. This means that the majority of bike trips occurred within the same region and very few trips occurred in more than one region.

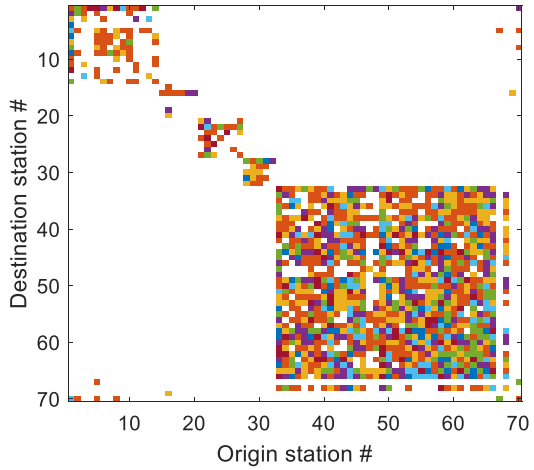

Figure 18. Adjacency matrix of the Bay Area BSS network.

Using PLSR as a regression algorithm makes it possible to build prediction models considering a multivariate response. Consequently, PLSR was applied to reduce the number of models to five, each of which is specified for one region (i.e., one ZIP code) to reflect the spatial correlation between stations. The input predictors' vector is X_t^z, which consists of the available bikes in each region z at time t, the month of year, day of week, time of day, and various selected weather conditions. The response's vector is $\log(Y_{t+\Delta}^z)$, where $z = 1, 2, 3, 4, 5$, which is the log of the number of available bikes at all stations in each of the studied regions z at a prediction horizon time Δ (ranges 15–120 minutes). We found that the prediction errors for PLSR were higher than the RF and LSBoost prediction errors when $\Delta = 15$ minutes, as shown in Figure 19. Although the prediction errors resulting from PLSR were higher than the previous results, the resulting models from PLSR were sufficient and desirable for relatively large BSS networks.

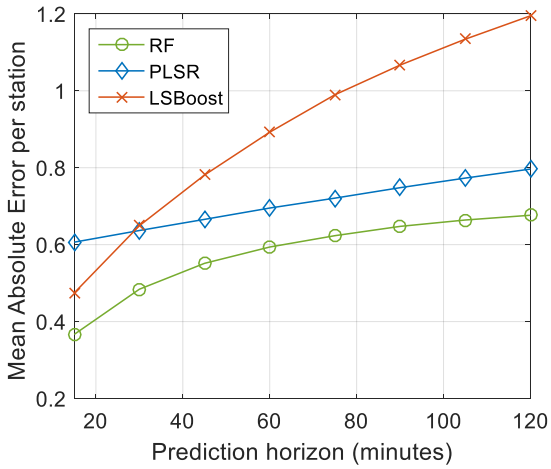

Figure 19. PLSR, RF, and LSBoost MAE at different prediction horizons.

5.7 Conclusions and Recommendations for Future Work

In this paper, we investigated modeling the number of available bikes at the San Francisco Bay Area BSS stations using machine learning algorithms. The investigation applied two approaches: 1) using RF and LSBoost univariate regression algorithms, and 2) using the PLSR multivariate regression algorithm. The univariate models were used to model the available bikes at each station. RF with a MAE of 0.37 bikes/station outperformed LSBoost with a MAE of 0.58 bikes/station. The multivariate model, PLSR, was applied to model available bikes at the spatially correlated stations of each region obtained from the trips' adjacency matrix. Results showed that the univariate models produced lower error predictions compared to the multivariate model, in which the MAE was approximately 0.6 bikes/station. However, the multivariate model results might be acceptable and reasonable when modeling the number of available bikes in BSS networks with a relatively large number of stations.

Investigating BSSs at the station-level is the full prediction problem that would help planners make decisions such as considering the relocation of underused stations (or building new ones) to serve busy areas in the network. The results from bike counts at stations located in downtown San Francisco were more volatile than the other stations in the BSS network. This high fluctuation in bike counts at these stations can be divided into two types: 1) fluctuation due to the periodic redistribution operation to rebalance bike counts; 2) fluctuation due to the high incoming/outgoing demand at the stations within a relatively short period of time. These findings suggest that demographic and built environment variables are critical influences in predicting bike

counts. Although the predictive model still lags by one step, the available bikes modeled at the station-level at time t had a noticeable influence on the prediction of bike counts at $t + \Delta$.

Investigating BSS networks in terms of determined regions gives new insights to policy makers. The fact that stations in each region derived by the multivariate analysis share the same ZIP code implies that most of the trips were short distance trips. This may be influenced by the overtime fees applied when trips are longer than 30 minutes. The results also illustrate that station neighbors and the prediction horizon time were found to be significant in modeling the number of available bikes. Specifically, when the prediction horizon time increases, the prediction error increases, with the most effective prediction horizon being 15 minutes. Determining prediction horizon is beneficial to policy makers and technicians to learn how to manage BSSs more responsively and achieve better performance in prediction.

5.8 Acknowledgment

This research effort was funded by the UrbComp program with a grant from the National Science Foundation.

References

[1] P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, p. 3, 2009.

[2] Bay Area Bike Share. (2016). *Introducing Bay Area Bike Share, your new regional transit system*. Available: http://www.bayareabikeshare.com/faq#BikeShare101

[3] C. Contardo, C. Morency, and L.-M. Rousseau, *Balancing a dynamic public bike-sharing system*. Cirrelt, 2012.

[4] J. Schuijbroek, R. Hampshire, and W.-J. van Hoeve, "Inventory rebalancing and vehicle routing in bike sharing systems," 2013.

[5] T. Raviv, M. Tzur, and I. A. Forma, "Static repositioning in a bike-sharing system: models and solution approaches," *EURO Journal on Transportation and Logistics*, vol. 2, no. 3, pp. 187-229, 2013// 2013.

[6] D. W. Daddio, "Maximizing Bicycle Sharing: an empirical analysis of capital bikeshare usage," 2012.

[7] X. Wang, G. Lindsey, J. E. Schoner, and A. Harrison, "Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations," *Journal of Urban Planning and Development*, vol. 142, no. 1, p. 04015001, 2015.

[8] C. Rudloff and B. Lackner, "Modeling demand for bicycle sharing systems–neighboring stations as a source for demand and a reason for structural breaks," 2013.

[9] R. Rixey, "Station-Level Forecasting of Bikesharing Ridership: Station Network Effects in Three US Systems," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2387, pp. 46-55, 2013.

[10] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns," *Procedia - Social and Behavioral Sciences*, vol. 20, pp. 514-523, 2011/01/01 2011.
[11] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," 2009, vol. 9, pp. 1420-1426.
[12] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455-466, 2010.
[13] J. W. Yoon, F. Pinelli, and F. Calabrese, "Cityride: a predictive bike sharing journey advisor," 2012, pp. 306-311: IEEE.
[14] C. Gallop, C. Tse, and J. Zhao, "A seasonal autoregressive model of Vancouver bicycle traffic using weather variables," *i-Manager's Journal on Civil Engineering*, vol. 1, no. 4, p. 9, 2011.
[15] Y.-C. Yin, C.-S. Lee, and Y.-P. Wong, "Demand Prediction of Bicycle Sharing Systems," ed: Stanford University.[Online], 2012.
[16] J. Du, R. He, and Z. Zhechev, "Forecasting Bike Rental Demand," ed: Stanford University, 2014.
[17] G. M. Dias, B. Bellalta, and S. Oechsner, "Predicting occupancy trends in Barcelona's bicycle service stations using open data," 2015, pp. 439-445: IEEE.
[18] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," 2015, p. 33: ACM.
[19] R. Giot and R. Cherrier, "Predicting bikeshare system usage up to one day ahead," 2014, pp. 22-29: IEEE.
[20] H. I. Ashqar, M. Elhenawy, M. H. Almannaa, A. Ghanem, H. A. Rakha, and L. House, "Modeling bike availability in a bike-sharing system using machine learning," in *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2017, pp. 374-378.
[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
[22] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14-23, 2011.
[23] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189-1232, 2001.
[24] Z. Barutçuoğlu and E. Alpaydın, "A comparison of model aggregation methods for regression," in *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003*: Springer, 2003, pp. 76-83.
[25] A. Höskuldsson, "PLS regression methods," *Journal of chemometrics*, vol. 2, no. 3, pp. 211-228, 1988.
[26] H. Wold, "Soft modelling: the basic design and some extensions," *Systems under indirect observation, Part II*, pp. 36-37, 1982.
[27] S. Wold, A. Ruhe, H. Wold, and I. W. Dunn, "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735-743, 1984.
[28] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109-130, 10/28/ 2001.

[29] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1-17, 1986/01/01 1986.
[30] B. Hamner. (2016). *SF Bay Area Bike Share | Kaggle*. Available: https://www.kaggle.com/benhamner/sf-bay-area-bike-share
[31] H. I. Ashqar, M. Elhenawy, A. Ghanem, M. H. Almannaa, and H. A. Rakha, "Modeling Bike Counts in a Bike-Sharing System Considering the Effect of Weather Conditions," 2016.
[32] Y. Ganjisaffar, R. Caruana, and C. V. Lopes, "Bagging gradient-boosted trees for high precision, low variance ranking models," 2011, pp. 85-94: ACM.
[33] World Population Review. (2017). *California Population 2017 (Demographics, Maps, Graphs)*. Available: http://worldpopulationreview.com/states/california-population/
[34] TOMTOM. (2017). *TomTom Traffic Index 2017*. Available: http://corporate.tomtom.com/releasedetail.cfm?ReleaseID=1012517
[35] Ford GoBike. (2017). *Station map*. Available: https://member.fordgobike.com/map/
[36] SF Station. (2017). *Harry Bridges Plaza*. Available: https://www.sfstation.com/harry-bridges-plaza-b7616

Chapter 6: Quality-Of-Service Measurement for BSS Stations

This chapter based on

Huthaifa I. Ashqar, Mohammed Elhenawy, Hesham A. Rakha, Leanna House. " A Proposed Quality-of-Service Measurement for Predicting Station Locations in Bike-Sharing Systems." *In review*, (2018)

6.1 Abstract

Bike-sharing systems (BSSs) are becoming an important part of urban mobility in many cities. BSSs are sustainable and environmentally friendly. Though, BSS operators spend great efforts to ensure bike and dock availability at each station. Measuring the quality-of-service (QoS) for each station or for the entire system has become an important issue for researchers. The traditionally-known QoS measurement reported in the literature is based on the proportion of *problematic stations*, which are defined as those with no bikes or docks available to users. It was found that it neither exposed the spatial dependencies between stations nor did it discriminate between stations in the BSS. Hence, we proposed a novel QoS measurement, namely *Optimal Occupancy*, in which: 1) The variations in arrival and pick up rates were considered in the formulation of Optimal Occupancy to capture the impact of heterogeneity in BSSs; 2) ANOVA analysis was used to prove it is discriminative; and 3) geo-statistics was applied to explore the spatial configuration of Optimal Occupancy variations and model variograms for spatial prediction. This study used anonymized bike trip dataset of 34 stations in downtown San Francisco to compare between the traditionally-known QoS measurement and Optimal Occupancy. Results revealed that the Optimal Occupancy is beneficial, outperforms the traditionally-known QoS measurement, and would result in better prediction of the QoS at nearby locations. In addition, Optimal Occupancy can also be used to predict candidate spots for the introduction of new stations in an existing BSS.

6.2 Introduction

In the next few decades, many traditional cities will be turned into smart cities, which are greener, safer, and faster. This transformation will be supported by recent advances in information and communication technology (ICT), the expected rapid spread of the Internet of Things (IoT),

and big data analytics. Smart cities may mitigate some of the negative impacts of traditional cities, which consume 75% of the world's energy and produce 80% of greenhouse gases [1]. Smart cities will rely on many components, including smart transportation. Smart transportation will integrate different transportation networks and allow them to work together so travelers and commuters can enjoy seamless multi-modal trips based on their preferences. Consequently, more commuters will be inspired to use public transportation systems, and many traffic-related problems, such as congestion, could be reduced.

A growing population, with more people living in cities, has led to increased pollution, noise, congestion, and greenhouse gas emissions. One possible approach to mitigating these problems is encouraging the use of bike-sharing systems (BSSs). BSSs are an integral part of urban mobility in many cities and are sustainable and environmentally friendly. As urban density increases, it is likely that more BSSs will appear due to their relatively low capital and operational costs, ease of installation, pedal assistance for people who are physically unable to pedal for long distances or on difficult terrain, and the ability to track bikes [2].

BSS operators take great efforts to ensure bike and dock availability at each station. This task can be difficult as the movement of users are highly dynamic, difficult to predict, and redistributing bikes is expensive. Recent studies have shown that there are spatial dependencies in bike usage at different stations [3-6], and that imbalances in the spatial distribution of bikes occur due to one-way use and short rental periods [6]. Thus, it is necessary for operators to understand the spatial dependencies to more effectively manage the system. For example, operators could improve the quality of service (QoS) by identifying the best candidate spots for new stations. However, finding the best QoS measurement for a station in a heterogeneous BSS and using it to study the spatial dependencies in the system is a challenging problem.

We investigated the state-of-art QoS measurement and found it to be largely indiscriminative at the station level. In this study, we propose a new QoS measurement, namely the *Optimal Occupancy*, to discriminate between different stations in heterogeneous BSSs. We demonstrate that Optimal Occupancy is not only discriminative but can also captures the spatial correlations in a BSS.

6.3 Related Work

Modeling bike sharing data is an area of significant research interest. In general, the main goals of previous studies have been to boost the redistribution operation [7-12], to gain new insights into and correlations between bike demand and other factors [13-17], and to support policy makers and managers in making optimized decisions [6, 13].

Research questions that have been studied previously include the strategic design, operation, and analysis of BSSs. Due to the potential benefits to operators, measuring the quality-of-service of stations or the entire system [18] has become an appealing issue for researchers. In some cases, operators measure the fraction of time that their stations are full or empty as a measurement of the QoS of the system [9]. Similarly, Fricker *et al.* considered the limiting probability that a station is empty or full as the performance measure. They argued that the optimal proportion of bikes at a station is slightly more than half the capacity of a station in a homogeneous system. In an heterogeneous system, however, they concluded that this performance metric collapses due to the heterogeneity [19].

Lin and Yang [20] investigated the strategic problems by studying the question of bike stations' measures of service. They argued that the measures of service quality in the system should include two measurements: the availability rate, which was defined as the proportion of pick-up requests at a bike station that are met by the bicycle stock on hand, and the coverage level, which is the fraction of the total demand at both origins and destinations that is within some specified time or distance from the nearest rental station. Fricker and Gast [21] proposed a stochastic model of a homogeneous BSS and investigated the impact of users' random choices on the number of problematic stations. Problematic stations were defined as stations that, at a given time, have no bikes available or no available docks for bikes to be returned to. Consequently, the performance of the system was determined by the proportion of problematic stations. However, these measures have critical drawbacks: (1) as BSSs usually offer two services: picking up bikes, and returning bikes; some measurements fail to take into account the QoS of returning bikes to stations; (2) some of the studies assume that, in contrast to real systems, the system is homogeneous; and (3) while some studies modeled the system as heterogeneous, they failed to consider the variability of the system parameters (i.e., arrival and pickup rates) throughout the same day or across the different days of the week and their dependency on the individual station.

In any BSS, one of the keys to success is the location and distribution of bike stations [20]. Some studies have worked on locating bike stations using different methods, such as location-allocation models [22, 23], and an optimization method that maximizes the demand covered and takes the available budget as a constraint [24]. The spatial distribution of the *potential demand* is a fundamental element in optimal location modeling. In order to estimate the potential demand, several studies used preference surveys to evaluate both the factors influencing the use of the bicycle mode and choice of routing [25-28]. Potential demand has also been estimated by considering the population, employment associated with each building, and the number of trips generated for each transport zone [22]. However, there are some limitations and drawbacks in the methods previously used to find the optimal station location: these methods are basically used to plan new systems and might not be useful to predict new stations in existing systems; they are aimed at serving the local population on selected days (e.g., workdays); and certain places in the studied area (e.g., large parks) have neither population nor jobs and yet may attract a considerable number of trips.

This paper makes two major contributions to the literature: (1) we propose a new discriminative QoS measure that reflects the spatial dependencies in a heterogeneous BSS which considers the variability of arrival and pickup rates; and (2) we use this QoS measure with geo-statistics to model a spatial variogram that could predict the QoS in nearby areas for the purpose of locating new stations in an existing BSS.

6.4 Proposed QoS Measurement

BSSs are highly heterogeneous; i.e. the arrival rates, pickup rates, origins, and destinations between stations in diverse areas and topographies are very different. These parameters may also vary with the time of day, day of the week, and season [29]. In this study, we consider that the bike-sharing system has N stations, in which each station i may have a unique capacity C_i (i.e., maximum number of docks). We assume that the dynamics of the system are as follows. Users reach the stations to pick up a bike at varying departure rates $\dot{D}_i$ at station i (we named it *departure rate* as their intention is to take the bike and depart to their destination stations). This departure rate $\dot{D}_i$ depends on station i and varies throughout the day and with different days of the week. If there are no available bikes, the user leaves the system or waits until another user arrives to return a bike. Users arrive at their destination stations to return the bike at a varying rate $\dot{A}_j$ at station j.

Similar to the departure rate, the arrival rate $\dot{A}_j$ depends on station j and varies throughout the day and with different days of the week. If there are less than C_j (i.e., capacity) bikes in this station, the user returns the bike and leaves the system. If the station is full, the user either chooses another station to return the bike or waits until another user reaches the station to pick up a bike.

To consider the impact of system heterogeneity, we introduce a new QoS measurement for each station, namely *Optimal Occupancy,* in which the variations of arrival and pick up rates are inherently in its definition. The Optimal Occupancy of a station is formulated in terms of two services: (1) picking up bikes, and (2) returning bikes. As each station i has a finite number of docks (i.e., capacity), two thresholds should be defined. The lower threshold (L_i) is the point when the number of available bikes ($B_{i,t}$) in station i at time t drops low enough that the possibility of a user not finding a bike is very high. The upper threshold (U_i) is the point when the number of bikes ($B_{i,t}$) in a station i at time t is high enough that the possibility of a user not finding a dock to return a bike is very high. For example, if a station's capacity is 25 docks and the number of available bikes at time t is within $[5, 20]$, then the station is considered functional; otherwise it needs to be rebalanced (i.e., it is a problematic station). In that sense, the Optimal Occupancy (O_{op}) is formulated as the ratio of the total time that a station is functional (t_f) during a given interval to the length of the interval (t_{total}):

$$O_{op_i} = \frac{t_{i,f}}{t_{i,total}} \tag{17}$$

where $t_{i,f} = \sum_{t=0}^{t=t_f} X_i(t)$ where $X_i(t)$ is the status function and defined as

$$X_i(t) = \begin{cases} 1, & \text{station } i \text{ is functional} \\ 0, & \text{station } i \text{ is problematic} \end{cases} \tag{18}$$

and station i is functional if $B_{i,t} \in [L_i, U_i]$ at any given time t. (19)

As the two thresholds L_i and U_i define the functionality of the station, L_i and U_i are correlated with the departure rate $\dot{D}_\iota$ and arrival rate $\dot{A}_\iota$, respectively. Both $\dot{D}_\iota$ and $\dot{A}_\iota$ randomly vary throughout the day, with different days of the week, and different months of the year. However, in this study, we assume that $\dot{D}_\iota$ and $\dot{A}_\iota$ vary only with different days of the week (DoW), and different months of the year (M) to be consistent with the length of the study interval (see Analysis and Results section). In fact, $\dot{D}_\iota$ and $\dot{A}_\iota$ are the bike counts picked up (D_i) or returned (A_i), respectively, per unit time ($t_{i,total}$). In that sense and to reflect the stochastic phenomenon

in the system, $\dot{D}_i$ and $\dot{A}_i$ were modeled using a Poisson Regression Model (PRM) with an exposure variable. Exposure is a measure of how the bike counts are divided. Since both rates are bike counts per unit time, time is considered as the exposure. The model contains a $log(t_{i,total})$, which is called the offset variable, as a term that could be added to the regression coefficients:

$$D_i \text{ or } A_i \sim Poisson\left(\theta_i^{(D) \text{ or } (A)}\right) \tag{20}$$

$$\text{where } \theta_i^{(D)} = \frac{\mu_i}{t_{i,total}}, \text{ and } \theta_i^{(A)} = \frac{\lambda_i}{t_{i,total}} \tag{21}$$

$$\log\left(\frac{\mu_i \text{ or } \lambda_i}{t_{i,total}}\right) = \beta_0 + \beta_1 DoW + \beta_2 M \tag{22}$$

$$L_i = \mu_i \tag{23}$$

$$U_i = C_i - \lambda_i \tag{24}$$

where μ and λ are the mean of the Poisson distribution for picked up bikes and returned bikes, respectively at each station i.

In that sense, problematic stations can be redefined as stations that, at any given time t, have fewer bikes available than the expected bike counts to be picked up during analysis discretization duration or more bikes than the difference between capacity and the expected bike counts to be returned during analysis discretization duration. The next sections in this study will further explain the concept of the proposed Optimal Occupancy QoS measurement by applying it to a real BSS dataset and comparing the new definition of problematic stations with the one previously used.

6.5 Dataset

One of the first BSSs in the United States was established in 1964 in Portland, with 60 bicycles available for public use. Although BSSs are still relatively limited, at present many cities, such as San Francisco and New York, have launched BSS programs. These programs implement different payment structures, conditions, and logistical strategies. In 2013, San Francisco launched the Bay Area Bike Share System (now called the "Ford GoBike" BSS), a membership-based system providing 24-hours-per-day, 7-days-per-week self-service access to short-term rental bicycles. Members can check out a bicycle from a network of automated stations, ride to the station

nearest their destination, and leave the bicycle safely locked for someone else to use [30]. The Bay Area BSS is designed for short, quick trips, and as a result, additional fees apply to trips longer than 30 minutes. In this system, bike stations connect users to transit, businesses, and other destinations in four areas: downtown San Francisco, Palo Alto, Mountain View, and downtown San Jose [30]. The Bay Area BSS is available to everyone 18 years and older with a credit or debit card. The system is designed to be used by commuters and tourists alike, whether they are trying to get across town at rush hour, traveling to and from the Bay Area Rapid Transit (BART) and Caltrain stations, or pursuing daily activities [30].

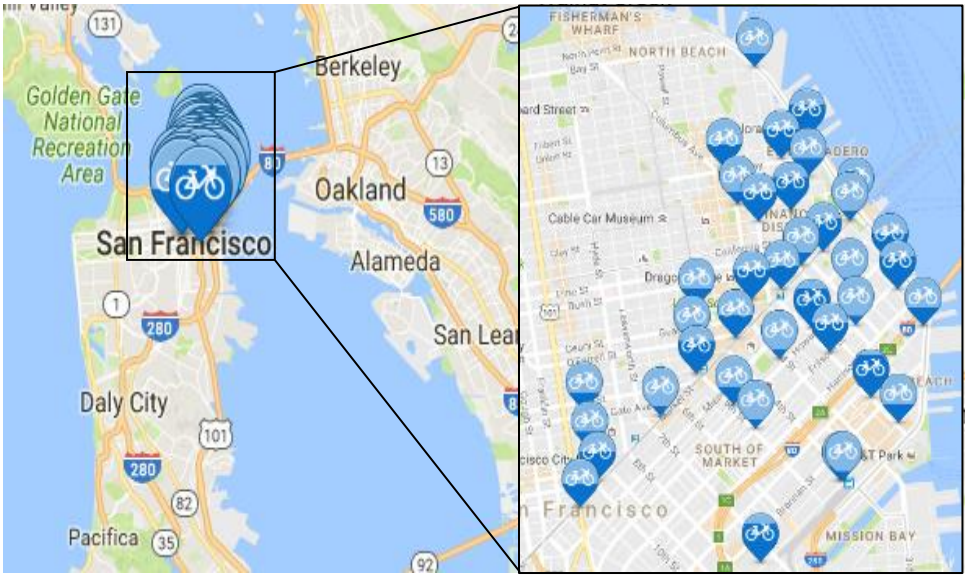

Figure 20. Stations map [30]

This study used anonymized bike trip data collected from August 2013 to August 2015 in San Francisco [31]. This study used two datasets of 34 stations in downtown San Francisco (Figure 20). The 34 stations have different capacities, ranging from 15 to 27 docks, which means the system is heterogeneous. The first dataset includes station ID, number of available bikes, number of available docks, and time of recording. The time data include the year, month, day of month, day of week, time of day, and minute at which a record was documented. As the database was updated every minute for 34 stations in San Francisco over 2 years, this dataset contains a large number of recorded incidents. The second dataset consists of the station ID, name of station, latitude and longitude of each station, the maximum number of docks, and the installation date. The latitude and longitude of each station were converted to the Universal Transverse Mercator Coordinate (UTM) system, which is expressed as a two-dimensional projection on the surface of

the Earth [32]. The UTM system divides the map of the Earth into 60 zones, each separated by 6 degrees in longitude. Locations are expressed in terms of Easting (i.e., the x coordinate) and Northing (i.e., the y coordinate). The UTM system provides coordinates on a worldwide flat grid that can be used to ease computation [33]. Figure 21 shows the conversion to the UTM system.

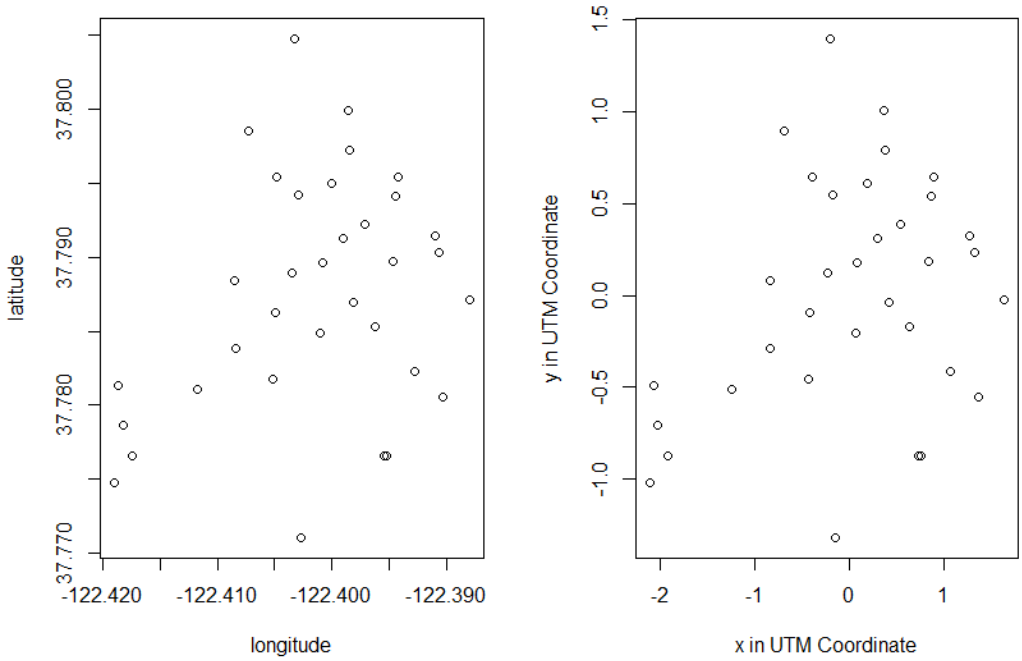

Figure 21. Conversion of latitude and longitude of each station to UTM system

6.6 Analysis and Results

In a BSS, the QoS measurement should reflect the spatial dependencies of BSS stations in addition to describing the performance of a station's service. Consequently, we investigated the traditionally-known QoS measurement using the Bay Area BSS dataset in San Francisco. We found that it was neither satisfying in exposing the spatial dependencies between stations nor adequate in describing the performance of the service.

The first QoS measurement presented in different studies (such as in [9, 19, 21]) is that of problematic stations, defined as stations that, at a given time, have no bikes available or no available spots for bikes to be returned to. This definition has been mainly used to describe the overall performance of the system. However, we used that definition to find a QoS measurement for a specific station by computing the ratio of the total time that a station is not problematic during

a given interval to the length of the interval. The second measurement is our proposed QoS measurement, Optimal Occupancy (O_{op}), which redefines problematic stations as stations that, at any given time t, have fewer bikes available than the expected bike counts to be picked up during analysis discretization duration or more bikes than the difference between capacity and the expected bike counts to be returned during analysis discretization duration. Similarly, we used our definition to find the Optimal Occupancy for a specific station by computing the ratio of the total time that a station is not problematic (i.e., functional) during a given interval to the length of the interval. For this specific dataset, and to effectively represent the service in the system, we defined the length of the study interval in both definitions as running from 8 a.m. to 5 p.m., which was found to be the peak hours for the system [34]. Figure 22 shows the locations of the stations with the corresponding results of the two QoS average measurements (over two years) of 34 stations in the Bay Area Bike Share in San Francisco. The measurements were first found for every 15 minutes at each station then averaged over the interval of the peak hours for the system.

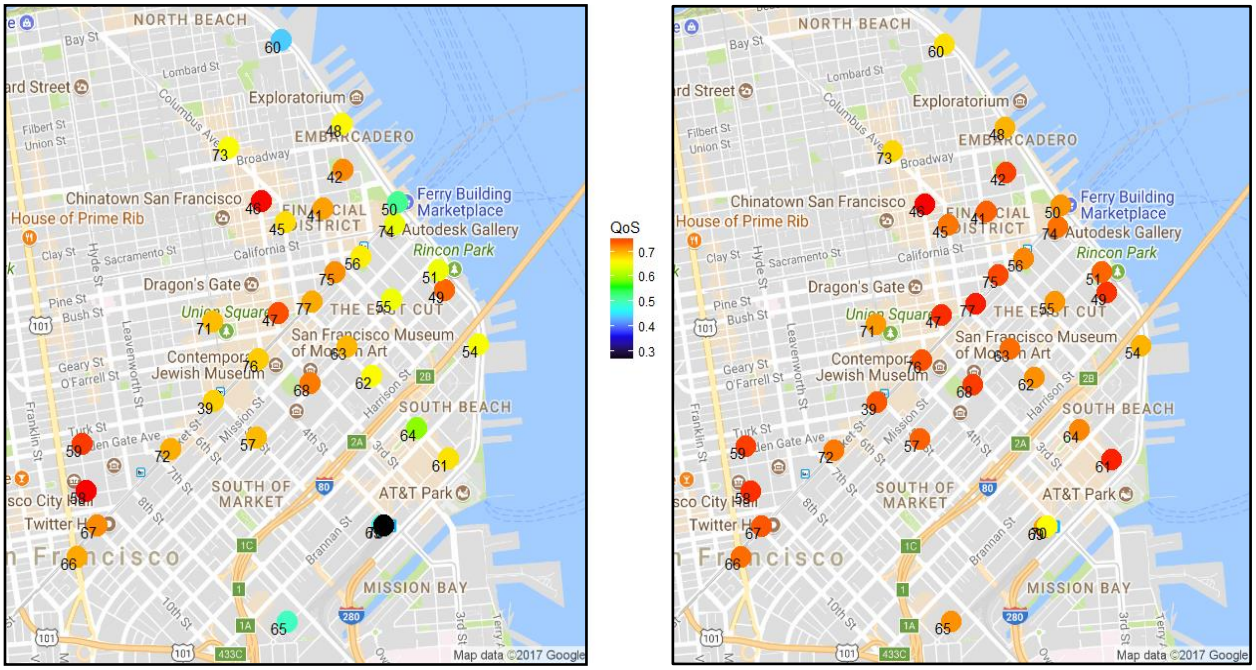

Figure 22. The locations, and the values of the (a) proposed QoS, and (b) traditionally-known QoS measurements

6.6.1 Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) was used to determine whether there are any statistically significant differences between the means of the two QoS measurements. Before interpreting the results of the hypothesis tests, we checked the ANOVA assumptions, and the hypothesis test results were found to be trusted. BSSs are highly heterogeneous, with arrival rates and pickup rates between stations in diverse areas and topographies varying with the time of day, day of the week, and season [29, 34]. Therefore, to fairly compare the two measurements, we compared the daily values for specific months and days. ANOVA was used to analyze the differences among four group means for all 34 stations: (1) Tuesdays of February, (2) Tuesdays of July, (3) Mondays of February, and (4) Mondays of July. The *p*-values resulting from testing the groups of traditionally-known QoS measurements are $0.7704, 0.8400, 0.5099,$ and 0.7443, respectively. This means that the Null Hypothesis is true and there are no significant differences ($p > 0.05$). On the other hand, the *p*-values resulting from testing the groups of the proposed QoS measurements (O_{op}) are $2.73E-29, 3.25E-36, 7.34E-41,$ and $1.42E-30$, respectively. This means that the Null Hypothesis is rejected and that there are significant differences between the measurements of the stations ($p < 0.05$). Figure 23 shows the differences among the Tuesdays of February group means for all 34 stations. It clearly demonstrates that the traditionally-known QoS cannot be used to discriminate between the stations, while the proposed Optimal Occupancy is discriminative to a sufficient extent. In that sense, recognition of the differences between the QoS of stations is not required in and of itself but because it is necessary for operators to effectively manage the system and it appears to reflect the dynamics of the BSS. Although we present the results of only four groups, in fact we examined the ANOVA test for other groups that cover most of the days of the week and months of the year. The results were found to be consistent with the results presented here.

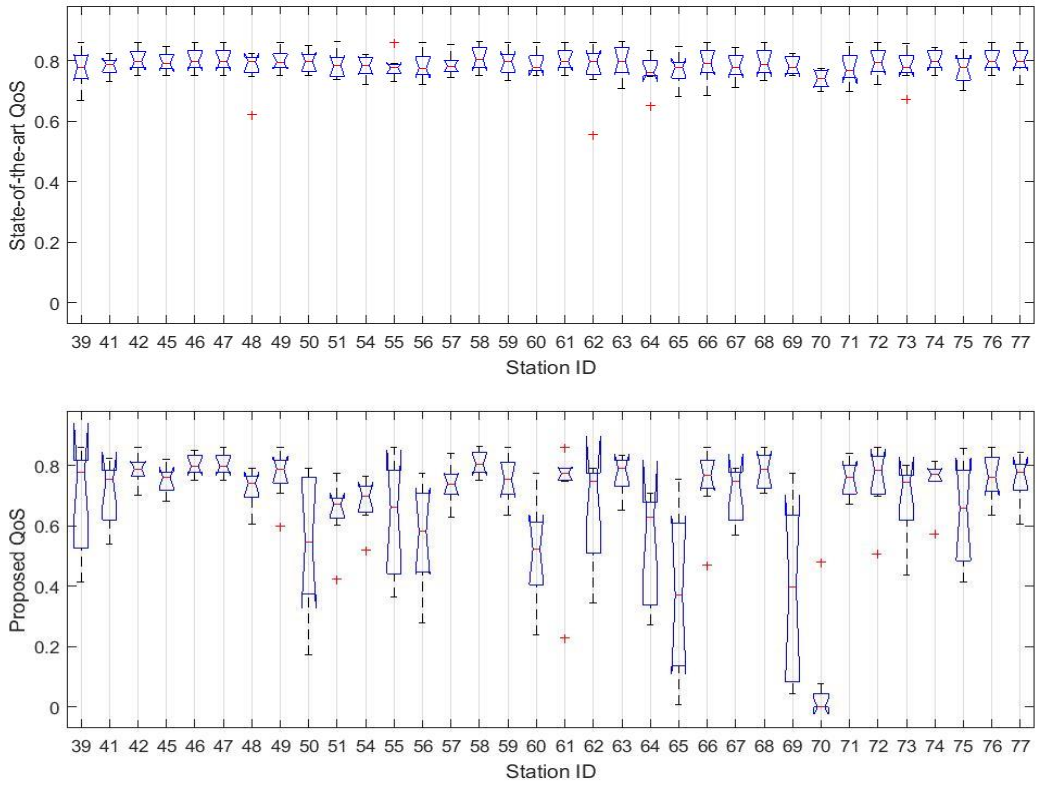

Figure 23. ANOVA test for Tuesdays of February for the 34 stations for (a) traditionally-known QoS, and (b) proposed QoS

6.6.2 Spatial Analysis

We applied geo-statistics to explore the spatial configuration of Optimal Occupancy variations. We used two packages in R, geoR to analyze geostatistical data [35] and gstat to perform geostatistical modelling and prediction [36]. The analysis was performed to assess whether the proposed Optimal Occupancy measurements can reflect the spatial dependencies and be used to predict the QoS in nearby areas. This will allow operators to determine candidate spots for new stations in the BSS, which will increase the overall QoS of the system.

Spatial statistics attempts to develop inferential methods to properly account for the spatial dependences in the presence of georeferenced observations. Spatial modeling typically contains a specification of a mean function and a model of the correlation structure (i.e. variogram), which is a description of the spatial continuity of the data. The variogram is the key function in geo-statistics as it is used to fit a model of the spatial correlation of the observed phenomenon [37, 38]. A variogram model is chosen by plotting the empirical variogram, which is a simple nonparametric

estimate of the variogram, and then comparing it to various theoretical shapes available. A variogram could be mathematically defined as [38]:

$$\gamma(\Delta x, \Delta y) = \frac{1}{2}\varepsilon\,[\{Z(x+\Delta x, y+\Delta y\} - Z(x,y)\}^2] \tag{25}$$

where $Z(x,y)$ is the value of the variable of interest at location (x,y), and ε [ ] is the statistical expectation operator. The variogram, $\gamma()$, is a function of the separation between points $(\Delta x, \Delta\, y)$, and not a function of the specific location (x,y). However, one common assumption of the spatial analysis is that it is isotropic. An isotropic variogram means that the correlation between any two observations depends only on the distance between those locations and not on their relative direction; otherwise, it is anisotropic [39].

A series of directional empirical variograms (including directions between 0° and 180°) was investigated to highlight the main observations' directions and check the spatial isotropy in the proposed QoS measurements data. The results illustrate that we cannot assume isotropy and that the directional empirical variogram for 45° outperforms other variograms as it reflects the correlation between the observations and the distance. The empirical variogram for 45° using transformed coordinates was estimated and is illustrated in Figure 24. It shows a steady increase in the semi-variance over increasing distance intervals to an absolute maximum between 1.0 and 1.5 km. For greater distances, Figure 24 displays an oscillatory state with a second maximum around 2.5 and 3 km.

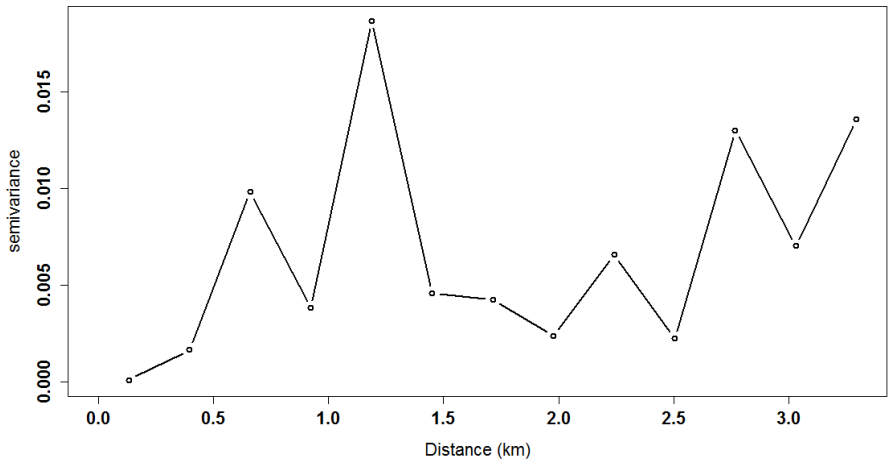

Figure 24. The empirical variogram for 45° using transformed coordinates

Modeling variograms are usually used for spatial prediction (i.e., interpolation). Most practical studies used Exponential, Spherical, and Gaussian models. As we assumed anisotropy, we applied the maximum likelihood estimation of spatial regression models to estimate the angle for geometric anisotropy of the three models. The Exponential variogram model yields the most beneficial realization of the spatial process in the BSS. While the Spherical model yields a decent estimation, the Gaussian model fails to fit a variogram that manifests the spatial correlation. We also applied the maximum likelihood estimation for the same three models to fit the traditionally-known QoS measurement to compare it with the Optimal Occupancy. Similarly, the Exponential variogram model outperforms the Spherical and the Gaussian models. Results in Table 14 show some inferences. According to the Bayesian information criterion (BIC) of the spatial and non-spatial models, the spatial model for Optimal Occupancy outperforms the non-spatial one, but the traditionally-known QoS non-spatial model outperforms the spatial one. This shows that the traditionally-known QoS cannot expose the spatial dependencies between stations. Therefore, using Optimal Occupancy is more advantageous than using the traditionally-known QoS. As the BIC for the spatial model demonstrates, Optimal Occupancy as a measurement is more gainful and would result in better prediction of the QoS in a BSS.

Table 14. Parameters estimation of the Exponential model for Optimal Occupancy and traditionally-known QoS

	BIC for spatial	BIC for non-spatial	Angle
Optimal Occupancy	-55.46	-52.12	78°
Traditionally-known QoS	-204.50	-211.10	71°

6.6.3 Optimal Location of New Stations

We proposed Optimal Occupancy as a QoS measurement to: (1) allow the operator to keep track of the performance of different stations in a BSS, so for example they may increase the number of docks/available bikes in a station; (2) identify the optimal location of new stations in existing systems using data-driven decision management approach. In the previous section, geo-statistics was used to model a spatial variogram that could predict the QoS in nearby areas for the purpose of locating new stations in an existing BSS. The model was used to produce new QoS datasets in order to build a QoS surface for the case study area. Figure 25 shows the QoS surface for the case study area in San Francisco. This surface could be used to quantify and visualize the QoS measurements represented by contours in the surface. Looking at the surface in Figure 25,

there are four hot spots (red-colored) that could be considered as candidates to add new stations nearby or increase the number of docks in a station. By doing so, we convert the surface into more homogeneous QoS terrain, which means BSS will be more functional (i.e. less problematic stations at any given time) and easier to rebalance. It is also interesting to mention that during our study, Ford GoBike, the operator of the case study BSS, has added different coming-soon stations near the abovementioned spots or added more docks to others. For example, a coming-soon station is to be built very near to Station 50 shown in Figure 22 (a), which we hypothesize it was added to increase the functionality of Station 50 (see [40]).

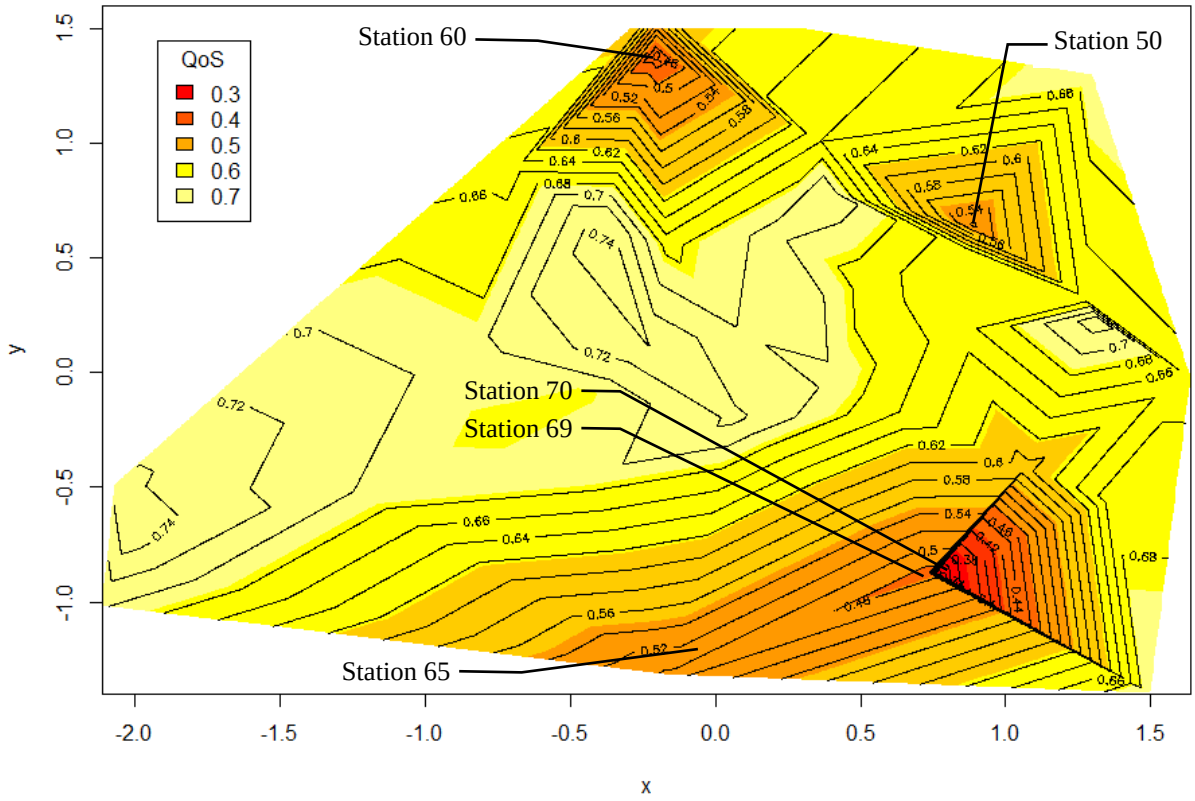

Figure 25. Predicted QoS surface for the case study area

In [41], a model was developed to predict the bike counts at each station in the Ford GoBike system using Random Forest (RF) as a univariate regression algorithm for different prediction horizons. Modeling bike counts using RF produced a Mean Absolute Error (MAE) of 0.37 bikes/station, which means the model was found to be promising. Station 50, namely Harry Bridges Plaza Station, was also found to be one of the stations that are highly unpredictable due to the high fluctuations in bike counts. When the area around Harry Bridges Plaza Station was

studied, it was hypothesized that this high incoming/outgoing demand comes from it being an open air area at the end of a market and restaurants, where artists, skaters, tourists and others congregate to enjoy the happenings and beautiful scenery [42]. In that sense, we will use the developed model in [41] to prove our hypothesis that if a new station is added, for example near Station 50, it will increase its functionality. We will compare the proposed QoS values for two different days of the week, Monday and Tuesday of July, before and after adding the new suggested station near Station 50. The model was used to predict the bike counts at Station 50 every 15 minutes for each of the selected days to estimate the proposed QoS using Equations (17) through (24). We assumed that the new station will cover only third of the two types of services that Station 50 used to serve. The resulting QoS for Station 50 was improved after adding the new suggested station by increasing from 0.52 to 0.84 and from 0.43 to 0.79 for Monday and Tuesday of July, respectively.

6.7 Conclusions

BSS operators tend to spend a great amount of time and effort to satisfy users. Accurately measuring the QoS of each station in a BSS will advance this mission. Moreover, measuring the QoS and using it to study the spatial dependencies in a BSS allows operators to better manage the system. For example, operators can determine candidate spots for new stations that will improve the overall QoS. Consequently, we investigated the traditionally-known QoS measurement and found it to be largely indiscriminative at the station level and not reflective of the spatial correlations. For that reason, we introduced a new QoS measurement, *Optimal Occupancy*. The Optimal Occupancy at a station is formulated in terms of two types of services: (1) picking up of bikes and (2) returning bikes. It is formulated as the ratio of the total time a station is functional during a given interval to the length of the interval. Consequently, we redefined problematic stations as stations that, at any given time, have fewer bikes available than the expected bike counts to be picked up during the analysis discretization duration or more bikes than the difference between capacity and the expected bike counts to be returned during the analysis discretization period.

We further studied the concept of the proposed QoS measurement by applying it to a real dataset of 34 stations in the San Francisco Area and also compared the new definition of problematic stations with the one previously used. First, results from ANOVA analysis clearly demonstrate that the traditionally-known QoS cannot be used to discriminate between the stations,

whereas the Optimal Occupancy is found to be sufficiently discriminative. Recognition of the differences between the QoS of stations benefits the effective management of the system and appears to reflect the dynamic nature of the BSS.

Second, we applied geo-statistics to explore the spatial configuration of the Optimal Occupancy variations and model variograms for spatial prediction. The empirical variogram shows a steady increase in semi-variance over increasing distance intervals to an absolute maximum between 1 and 1.5 km. The Exponential variogram model was fitted and yields the most beneficial realization of the spatial process in the BSS. Results revealed that the spatial model for Optimal Occupancy outperforms the non-spatial one. Furthermore, Optimal Occupancy as a measurement is more gainful and would result in better prediction for the QoS in nearby locations. However, the spatial model was used to produce new QoS datasets in order to build a QoS surface for the case study area. Adding new stations nearby the hot spots in the surface, we could convert the surface into a more homogeneous QoS terrain, indicating that the BSS will be more functional and easier to rebalance as a result of this change. For example, the resulting QoS for Station 50 was improved after adding the new suggested station from 0.52 to 0.84 and from 0.43 to 0.79 for Monday and Tuesday of July, respectively.

6.8 Acknowledgements

This effort was funded by the Urban Mobility and Equitable Center and the NSF UrbComp project.

References

[1] S. P. Mohanty, U. Choppali, and E. Kougianos, "Everything you wanted to know about smart cities: The internet of things is the backbone," *IEEE Consumer Electronics Magazine*, vol. 5, no. 3, pp. 60-70, 2016.

[2] P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, p. 3, 2009.

[3] P. Borgnat, E. Fleury, C. Robardet, and A. Scherrer, "Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program," in *ECCS'09*, Warwick, United Kingdom, 2009: Complex Systems Society.

[4] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," 2009, vol. 9, pp. 1420-1426.

[5] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455-466, 2010.

[6] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns," *Procedia - Social and Behavioral Sciences,* vol. 20, pp. 514-523, 2011/01/01 2011.

[7] L. Caggiani, R. Camporeale, M. Ottomanelli, and W. Y. Szeto, "A modeling framework for the dynamic management of free-floating bike-sharing systems," *Transportation Research Part C: Emerging Technologies,* vol. 87, pp. 159-182, 2018/02/01/ 2018.

[8] C. Contardo, C. Morency, and L.-M. Rousseau, *Balancing a dynamic public bike-sharing system*. Cirrelt, 2012.

[9] J. Schuijbroek, R. Hampshire, and W.-J. van Hoeve, "Inventory rebalancing and vehicle routing in bike sharing systems," 2013.

[10] T. Raviv, M. Tzur, and I. A. Forma, "Static repositioning in a bike-sharing system: models and solution approaches," *EURO Journal on Transportation and Logistics,* vol. 2, no. 3, pp. 187-229, 2013// 2013.

[11] Y. Liu, W. Y. Szeto, and S. C. Ho, "A static free-floating bike repositioning problem with multiple heterogeneous vehicles, multiple depots, and multiple visits," *Transportation Research Part C: Emerging Technologies,* vol. 92, pp. 208-242, 2018/07/01/ 2018.

[12] A. Pal and Y. Zhang, "Free-floating bike sharing: Solving real-life large-scale static rebalancing problems," *Transportation Research Part C: Emerging Technologies,* vol. 80, pp. 92-116, 2017/07/01/ 2017.

[13] D. W. Daddio, "Maximizing Bicycle Sharing: an empirical analysis of capital bikeshare usage," 2012.

[14] X. Wang, G. Lindsey, J. E. Schoner, and A. Harrison, "Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations," *Journal of Urban Planning and Development,* vol. 142, no. 1, p. 04015001, 2015.

[15] C. Rudloff and B. Lackner, "Modeling demand for bicycle sharing systems–neighboring stations as a source for demand and a reason for structural breaks," 2013.

[16] R. Rixey, "Station-Level Forecasting of Bikesharing Ridership: Station Network Effects in Three US Systems," *Transportation Research Record: Journal of the Transportation Research Board,* no. 2387, pp. 46-55, 2013.

[17] M. Bordagaray, L. dell'Olio, A. Fonzone, and Á. Ibeas, "Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques," *Transportation Research Part C: Emerging Technologies,* vol. 71, pp. 231-248, 2016/10/01/ 2016.

[18] A. Gunasekaran, C. Patel, and E. Tirtiroglu, "Performance measures and metrics in a supply chain environment," *International journal of operations & production Management,* vol. 21, no. 1/2, pp. 71-87, 2001.

[19] C. Fricker, N. Gast, and H. Mohamed, "Mean field analysis for inhomogeneous bike sharing systems," 2012.

[20] J.-R. Lin and T.-H. Yang, "Strategic design of public bicycle sharing systems with service level constraints," *Transportation research part E: logistics and transportation review,* vol. 47, no. 2, pp. 284-294, 2011.

[21] C. Fricker and N. Gast, "Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity," *Euro journal on transportation and logistics,* vol. 5, no. 3, pp. 261-291, 2016.

[22] J. C. García-Palomares, J. Gutiérrez, and M. Latorre, "Optimizing the location of stations in bike-sharing programs: A GIS approach," *Applied Geography*, vol. 35, no. 1, pp. 235-246, 2012/11/01/ 2012.
[23] G. Rybarczyk and C. Wu, "Bicycle facility planning using GIS and multi-criteria decision analysis," *Applied Geography*, vol. 30, no. 2, pp. 282-293, 2010.
[24] I. Frade and A. Ribeiro, "Bike-sharing stations: A maximal covering location approach," *Transportation Research Part A: Policy and Practice*, vol. 82, no. Supplement C, pp. 216-227, 2015/12/01/ 2015.
[25] J. Dill and K. Voros, "Factors affecting bicycling demand: initial survey findings from the Portland, Oregon, region," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2031, pp. 9-17, 2007.
[26] J. E. Abraham, S. McMillan, A. T. Brownlee, and J. D. Hunt, "Investigation of cycling sensitivities," 2002.
[27] K. Shafizadeh and D. Niemeier, "Bicycle journey-to-work: travel behavior characteristics and spatial attributes," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1578, pp. 84-90, 1997.
[28] L. D. O. Meng, "Implementing bike-sharing systems," *Proceedings of the Institution of Civil Engineers*, vol. 164, no. 2, p. 89, 2011.
[29] P. Borgnat, P. Abry, P. Flandrin, C. Robardet, J.-B. Rouquier, and E. Fleury, "Shared bicycles in a city: A signal processing and data analysis perspective," *Advances in Complex Systems*, vol. 14, no. 03, pp. 415-438, 2011.
[30] Bay Area Bike Share. (2016). *Introducing Bay Area Bike Share, your new regional transit system*. Available: http://www.bayareabikeshare.com/faq#BikeShare101
[31] B. Hamner. (2016). *SF Bay Area Bike Share | Kaggle*. Available: https://www.kaggle.com/benhamner/sf-bay-area-bike-share
[32] National Geodetic Survey. (2017). *Universal Transverse Mercator Coordinates*. Available: https://geodesy.noaa.gov/TOOLS/utm.shtml#
[33] Geokov. (2014). *UTM - Universal Transverse Mercator*. Available: http://geokov.com/education/utm.aspx
[34] M. H. Almannaa, M. Elhenawy, A. Ghanem, H. I. Ashqar, and H. A. Rakha, "Network-wide bike availability clustering using the college admission algorithm: A case study of San Francisco Bay area," in *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2017, pp. 580-585.
[35] P. J. Ribeiro Jr and P. J. Diggle, "geoR: a package for geostatistical analysis," *R news*, vol. 1, no. 2, pp. 14-18, 2001.
[36] E. J. Pebesma, "Multivariable geostatistics in S: the gstat package," *Computers & Geosciences*, vol. 30, no. 7, pp. 683-691, 2004.
[37] N. Cressie, *Statistics for spatial data*. John Wiley & Sons, 2015.
[38] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.
[39] A. Maity and M. Sherman, "Testing for Spatial Isotropy Under General Designs," *Journal of statistical planning and inference*, vol. 142, no. 5, pp. 1081-1091, 2012.
[40] Ford GoBike. (2018). *Station map*. Available: https://member.fordgobike.com/map/
[41] H. I. Ashqar, M. Elhenawy, and H. A. Rakha, "Network and Station-Level Bike-Sharing System Prediction: A San Francisco Bay Area Case Study," *unpublished*, 2018.

[42] SF Station. (2017). *Harry Bridges Plaza*. Available: https://www.sfstation.com/harry-bridges-plaza-b7616

Chapter 7: Conclusions and Future Research

Due to relatively low capital and operational costs, as well as ease of installation, many cities in the U.S. are making investments in BSSs. BSSs suffer from several planning problems that could be divided into three levels; a strategic, a tactical, and an operational level. This dissertation is a building block for a smart BSS in the strategic level, which could be used in real and different applications. Generally, four components were developed in the dissertation: transportation mode recognition, quantifying the effect of various features on BSS, network and station-level BSS prediction, and quality-of-service measurement for BSS stations.

This dissertation presented the development of four components in order to strategically design a smart BSS for smart city. Results of the first component show that the classification accuracy of the proposed framework outperforms traditional approaches. Transforming the time domain features to the frequency domain also adds new features in a new space and provides more control on the loss of information. Consequently, combining the time domain and the frequency domain features in a large pool and then choosing the best subset results in higher accuracy than using either domain alone. The proposed two-layer classifier obtained a maximum classification accuracy of 97.02%. It was also demonstrated that the proposed approach in the second component is promising to quantify the effect of various features on a large BSS and on each station in cases of large networks with big data. The results show that the time-of-the-day, temperature, and humidity level are significant count predictors. It also shows that as weather variables are geographic location dependent and thus should be quantified before using them in modeling. Results of the third component show that univariate models had lower error predictions than the multivariate model. Moreover, results of the station-level analysis suggested that demographic and built environment variables were critical factors in predicting bike counts. We also demonstrate that the available bikes modeled at the station-level at time t had a notable influence on the bike count models. The multivariate model results will be reasonable at the network-level, with a relatively large number of spatially correlated stations. Results of the fourth component revealed that Optimal Occupancy as a measurement is beneficial and would result in better prediction for the QoS at nearby locations. In addition, Optimal Occupancy can also be used to predict candidate spots for new stations in a BSS.

For the future research, the presented methods, models, and measurement might be investigated more using different datasets from other BSSs. This should reveal what aspects and factors that were not included in the proposed methods, models, and measurement. All this work might be correlated and tested in the context of operator policies and regulations. One might also investigate the applicability of these proposed methods, models, and measurement on different BSS schemes such as the free-floating scheme.