Analysis of Free-floating Bike Sharing and Insights on System Operations

or

Analyzing Mobility Patterns and Imbalance of Free Floating Bike Sharing Systems

Center for Transportation, Environment, and Community Health
Final Report

by

Aritra Pal, Yu Zhang, Changhyun Kwon

January 31, 2018

DISCLAIMER

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2.Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Analysis of Free-floating Bike Sharing and Insights on System Operations *or* Analyzing Mobility Patterns and Imbalance of Free Floating Bike Sharing Systems		5. Report Date January 31, 2018 6. Performing Organization Code
7. Author(s) Aritra Pal Yu Zhang (ORCID ID 0000-0003-1202-626X) Changhyun Kwon (ORCID ID 0000-0001-8455-6396)		8. Performing Organization Report No.
9. Performing Organization Name and Address Civil and Environmental Engineering University of South Florida Tampa, FL 33620		10. Work Unit No. 11. Contract or Grant No. 69A3551747119
12. Sponsoring Agency Name and Address U.S. Department of Transportation 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Final Report 11/30/2016-11/29/2017 14. Sponsoring Agency Code US-DOT
15. Supplementary Notes		

16. Abstract

Bike Sharing is a sustainable mode of urban mobility, not only for regular commuters but also for casual users and tourists. Freefloating bike sharing (FFBS) is an innovative bike sharing model, that saves on start-up cost, prevents bike theft, and offers significant opportunities for smart management by tracking bikes in real-time with built-in GPS. The primary objective of this paper is to understand the mobility patterns and imbalance of an FFBS by analyzing its historical trip and weather data. Resulting outcomes provide insights to assist the system operator to make more informed decisions. Researchers have studied mobility patterns by analyzing historical trip and weather data of station-based bike sharing systems (SBBS) using data visualization and or generalized linear models. However, none of these studies considered interaction between independent variables or study imbalance as a dependent variable. In this paper, we demonstrate that by considering such interactions, more insights can be obtained about the mobility patterns and imbalance of an FFBS. We propose a simple method to decompose continuous variables into binary variables and two stage models that consider interactions between independent variables. The proposed decomposition method significantly improves the (quasi-)Poisson regression model commonly used in the literature and has the ability to identify intervals of a continuous variable for which they are statistically significant.

17. Key Words	18. Distribution Statement		
Sharing mobility, Statistical modeling, econometrics modeling, Poisson regression, bike sharing operational management	Public Access		
19. Security Classif (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No of Pages	22. Price

Form DOT F 1700.7 (8-69)

Analyzing Mobility Patterns and Imbalance of Free Floating Bike Sharing Systems

Aritra Pal*[1], Yu Zhang†[2, 3], and Changhyun Kwon‡[1]

[1]Department of Industrial and Management Systems Engineering, University of South Florida
[2]Department of Civil and Environmental Engineering, University of South Florida
[3]College of Transportation Engineering, Tongji University, China

Abstract

Bike Sharing is a sustainable mode of urban mobility, not only for regular commuters but also for casual users and tourists. Free-floating bike sharing (FFBS) is an innovative bike sharing model, that saves on start-up cost, prevents bike theft, and offers significant opportunities for smart management by tracking bikes in real-time with built-in GPS. The primary objective of this paper is to understand the mobility patterns and imbalance of an FFBS by analyzing its historical trip and weather data. Resulting outcomes provide insights to assist the system operator to make more informed decisions. Researchers have studied mobility patterns by analyzing historical trip and weather data of station-based bike sharing systems (SBBS) using data visualization and or generalized linear models. However, none of these studies considered interaction between independent variables or study imbalance as a dependent variable. In this paper, we demonstrate that by considering such interactions, more insights can be obtained about the mobility patterns and imbalance of an FFBS. We propose a simple method to decompose continuous variables into binary variables and two stage models that consider interactions between independent variables. The proposed decomposition method significantly improves the (quasi-)Poisson regression model commonly used in the literature and has the ability to identify intervals of a continuous variable for which they are statistically significant.

Keywords: Free-floating bike sharing; quantiles; interactions; regularization; negative binomial regression;

Introduction

Free-floating bike sharing (FFBS), also known as station-less bike sharing, is a new generation of bike sharing systems (BSS) that allows bikes to be locked to ordinary bike racks (or any solid frame

*aritra1@mail.usf.edu
†Corresponding Author:yuzhang@usf.edu
‡chkwon@usf.edu

or standalone), eliminating the need for specific stations. It saves on start-up cost by avoiding the construction of expensive docking stations and kiosk machines required for station-based bike sharing (SBBS). With built-in GPS, customers can find and reserve bikes via a smart phone or a web app, and operators can track the usage of the bikes in real-time. Such systems have two primary benefits. First, user satisfaction levels increase, as renting and returning bikes become extremely convenient, and second, operators have a basis for smart management of the system. For historical information on BSS and a more detailed comparison between FFBS and SBBS, refer to DeMaio [13] and Pal and Zhang [25] respectively.

In the case of SBBS, the core problem faced by operators is maximizing the service level by maintaining an optimal inventory of bikes at each station, because excess supply may hamper the return of bikes, whereas shortage in supply may result in increased access cost for users (e.g., elongated walking distance) or in lost demand. FFBS has two prevalent models for parking bikes. In one, designated parking areas (physical or geo-fencing) are provided in public space with or without bike racks, and in the other, bikes are allowed to be parked at any legal parking sites, i.e., sites without violating the right of way. The first model leads to a system very similar to station-based but with a much larger number of parking areas, because the cost of constructing those designated parking areas, even with bike racks, is less than one tenth the cost of constructing docking stations. The second model has quite different features. Bikes could be scatted all over the service region. For this model, the return of bikes is not an issue, but the imbalance of demand and supply will result in lost demand if at a particular zone (defined by the radius of willingness to walk), demand is higher than supply. Also, it is possible that operators employ a hybrid model, i.e., allowing bikes to be parked in designated parking areas for some zones but any legal parking sites in other zones. To mitigate the overall or a station/zonal imbalance, the operator may use different types of rebalancing strategies depending on the situation at hand. For a more detailed description of various rebalancing strategies available to operators, refer to Pal and Zhang [25].

Solving the core problem of an established BSS requires the understanding of the mobility patterns of its users. It enables the operator to estimate an approximate target distribution of bikes for rebalancing as well as gain insights necessary for developing appropriate rebalancing strategies by addressing issues such as whether static rebalancing is sufficient or dynamic rebalancing is needed, when the different types of rebalancing should start, and how much time is available for each type of rebalancing. In this paper, we demonstrate our proposed methods of understanding mobility patterns and extracting management insights, using the historical trip data of Share-A-Bull BSS (SABB), an FFBS on the Tampa campus of the University of South Florida (USF). The knowledge and insights gained using our proposed method can be used by operators of both FFBS and SBBS to improve their respective service levels.

Existing studies on mobility patterns analysis focus primarily on SBBS by analyzing historical trip and weather data. Authors take system outputs (rentals and or returns) as dependent variables and environmental factors, socio-demographic features and cycling infrastructure as independent variables. However, none of these studies, consider imbalance (difference between returns and rentals) as a dependent variable or *interaction* between the independent variables. In this paper, we demonstrate that by considering imbalance as a dependent variable and the *interaction* between independent variables, more knowledge and insights can be obtained about the mobility patterns of an FFBS, than by using conventional methods like data visualization and generalized linear models.

To be consistent with other studies in the literature, rentals and returns of a BSS are referred to as pickups and dropoffs respectively, in the rest of the paper. To be more specific, in this paper, we are trying to determine how the demand (dropoffs and pickups) and imbalance of an FFBS vary with time and how they are affected by exogenous variables such as holidays, weather conditions, etc. To accomplish this, we propose a simple method to decompose continuous variables into binary variables that improves the base model (Poisson and negative binomial regression models) commonly used in the literature as well as consider all feasible (second and third order) interactions between binary variables. The purpose of adding such interactions is to extract additional insights from the data for operational management purposes. It is obvious that considering *interactions* could result in a significant increase in the number of independent variables, sometimes even significantly larger than the number of observations. This makes it inappropriate to use (generalized) linear models directly. To address this issue, we first use a regularization operator to shrink the variable space and then estimate an appropriate linear model on the shrunk variable space. Although our case study is an FFBS, our proposed method can be used for SBBS without any modifications.

The remainder of the paper is organized as follows. Section 2 summarizes and highlights gaps in the literature. Section 3 describes the proposed method. Section 4 introduces the case study and presents the experimental results of our proposed methods. Section discusses how knowledge and operational management insights about the SABB FFBS can be drawn from the statistical models. We also demonstrates, how some of this insights can be used for making useful recommendations to the operator of the system. Finally, Section 6 concludes the paper with directions for future research.

Literature Review

Papers related to analytics of a BSS (primarily SBBS) can be broadly classified into two categories, based on their objective(s): 1) papers whose primary objective is to predict the future demand of the system and 2) papers whose primary objective is to understand and describe a system(s), so that either its service level can be improved or the system can be expanded. The most important papers related to predicting the future demand of a BSS (or car sharing systems) are Borgnat et al. [8], Cheu et al. [10], Kaltenbrunner et al. [23], Regue and Recker [26] and Alvarez-Valdes et al. [7]. It is interesting to note that, papers focused on predicting future demand almost always rely on non-parametric statistical methods, like neural networks (Cheu et al. [10]), gradient boosted machines (Regue and Recker [26]), non-homogeneous Poisson process (Alvarez-Valdes et al. [7]), etc. Further, recent papers on predicting demand (Alvarez-Valdes et al. [7], Regue and Recker [26]) also use the outputs of their demand prediction model as inputs to a rebalancing optimization model.

On the other hand, papers in the second category always use generalized linear and generalized linear mixed models as their core statistical method. This is because linear models are easy to interpret compared to non-linear and non-parametric models. Papers in the second category can be further subdivided into two subcategories: 1) papers that try to understand factors affecting the demand of a BSS and 2) papers that propose metrics either to compare several BSS among themselves or to measure the performance of a BSS. In the first subcategory, the most common factors considered in the literature are:

1. temporal factors (season, month, day of week, holiday and hour of day) - Faghih-Imani and Eluru [15], Faghih-Imani et al. [16], Gebhart and Noland [21], Wagner et al. [28]

2. meteorological factors (temperature, relative humidity, wind speed, etc) - Caulfield et al. [9], Faghih-Imani and Eluru [15], Faghih-Imani et al. [16], Gebhart and Noland [21]

3. socio-demographic factors - Faghih-Imani et al. [16, 17]

4. infrastructure of BSS and other modes of transportation - Faghih-Imani et al. [14], Faghih-Imani and Eluru [15], Faghih-Imani et al. [16, 17]

5. size of operating area (large, medium or small-scale city) - Caulfield et al. [9]

6. effect of expansion on demand - Wagner et al. [28], Zhang et al. [30]

.

Contrary to the above mentioned papers, Fishman et al. [18] studied factors that affect membership instead of demand of a BSS. In the second subcategory, papers such as de Chardon and Caruso [11], de Chardon et al. [12], OBrien et al. [24] propose methods to compare several BSS using daily trip data, whereas de Chardon and Caruso [11], de Chardon et al. [12] propose metrics to measure the quality and performance of a BSS without using the daily trip data.

To the best of our knowledge, none of the papers in the literature, consider imbalance as a dependent variable or interactions between independent variables. Thus, this is the first paper on an FFBS, which takes imbalance as a dependent variable and considers *interactions between independent variables* in a statistical model. We propose two stage models to address the increase in the number of independent variables when *interactions between independent variables* are considered. Although in this paper, we are focused on extracting knowledge and insight, often smart use of interactions between independent variables can lead to significant improvement in prediction accuracy (or decrease in out of sample testing error). We also propose a simple method to decompose continuous variables into binary variables, which significantly improves the negative binomial regression model commonly used in the literature, and has the ability to identify intervals of a continuous variable that are statistically significant. Further, our proposed methodology provides an unique opportunity to study an FFBS and make recommendations to the operator from various vantage points.

Table 1: Dependent variables used in this paper

Variable Name	Variable Description
Daily Dropoffs	Number of dropoffs in that day
Hourly Dropoffs	Number of dropoffs in that hour
Daily Pickups	Number of pickups in that day
Hourly Pickups	Number of pickups in that hour
Imbalance	Difference of the number of dropoffs and pickups in that hour

Methodology

In this section, we describe the variables used in this paper, method of collecting and cleaning the data, strategy for discretizing continuous variables into binary variables, method for creating interactions between independent binary variables, and two stage models for scenarios when number of independent variables outnumbers number of observations.

Variables

In this paper, the dependent variables are daily and hourly dropoffs and pickups as well as hourly imbalance. Hourly imbalance equals the difference of the number of dropoffs and the number of pickups in that hour. Unlike dropoffs and pickups, we do not study daily imbalance as its mean and variance is zero and close to zero respectively. This makes perfect sense, as the daily dropoffs and pickups will be close to each other unless bikes are added to or removed from the system by the operator. Daily and hourly dropoffs and pickups are non-negative count variables whereas hourly imbalance is a variable which can take any value from the set of real numbers.

Independent variables used in this paper include temporal variables (season, month, day and hour) and holiday and weather variables (temperature, apparent temperature, relative humidity, wind speed, cloud cover and dew point). Season, month and day are nominal variables whereas hour is an ordinal variable. To have correct estimates, we decompose both nominal and ordinal variables in to binary (or dummy) variables for each level. Holiday is a binary variable and the six weather variables are continuous. Tables 1, 2 and 3 provide a more detailed description of the dependent variables, binary independent variables and continuous independent variables respectively.

Data descriptions

We test our proposed methods on the SABB FFBS program at USF, Tampa. Phase I of the program was launched in August 2015, providing 100 bikes to students, staff and faculty at no charge if the users limited their cumulative usage time to less than two hours per day. An hourly fee was imposed for the extra time beyond the daily two hour free quota. With Phases II and III in the coming years, the program will be expanded to 300 bikes and cover both the Tampa campus and student housing in the vicinity of the campus. The program is expected to be integrated with parking management and other multi-modal transportation initiatives on the campus. USF

Table 2: Binary independent variables used in this paper

Variable Name	Variable Description
Spring Season Indicator	1 if Spring, 0 otherwise
Autumn Season Indicator	1 if Autumn, 0 otherwise
Summer Season Indicator	1 if Summer, 0 otherwise
Fall Season Indicator	1 if Fall , 0 otherwise
January Indicator	1 if January, 0 otherwise
February Indicator	1 if February, 0 otherwise
March Indicator	1 if March, 0 otherwise
April Indicator	1 if April, 0 otherwise
May Indicator	1 if May, 0 otherwise
June Indicator	1 if June, 0 otherwise
July Indicator	1 if July, 0 otherwise
August Indicator	1 if August, 0 otherwise
September Indicator	1 if September, 0 otherwise
October Indicator	1 if October, 0 otherwise
November Indicator	1 if November, 0 otherwise
December Indicator	1 if December, 0 otherwise
Monday Indicator	1 if Monday, 0 otherwise
Tuesday Indicator	1 if Tuesday, 0 otherwise
Wednesday Indicator	1 if Wednesday, 0 otherwise
Thursday Indicator	1 if Thursday, 0 otherwise
Friday Indicator	1 if Friday, 0 otherwise
Saturday Indicator	1 if Saturday, 0 otherwise
Sunday Indicator	1 if Sunday, 0 otherwise
Holiday Indicator	1 if Saturday or Sunday or a US Holiday, 0 otherwise
Hour 0 Indicator (00:00)	1 if after 12:00 AM and before 1:00 AM, 0 otherwise
Hour 1 Indicator (01:00)	1 if after 1:00 AM and before 2:00 AM, 0 otherwise
Hour 2 Indicator (02:00)	1 if after 2:00 AM and before 3:00 AM, 0 otherwise
Hour 3 Indicator (03:00)	1 if after 3:00 AM and before 4:00 AM, 0 otherwise
Hour 4 Indicator (04:00)	1 if after 4:00 AM and before 5:00 AM, 0 otherwise
Hour 5 Indicator (05:00)	1 if after 5:00 AM and before 6:00 AM, 0 otherwise
Hour 6 Indicator (06:00)	1 if after 6:00 AM and before 7:00 AM, 0 otherwise
Hour 7 Indicator (07:00)	1 if after 7:00 AM and before 8:00 AM, 0 otherwise
Hour 8 Indicator (08:00)	1 if after 8:00 AM and before 9:00 AM, 0 otherwise
Hour 9 Indicator (09:00)	1 if after 9:00 AM and before 10:00 AM, 0 otherwise
Hour 10 Indicator (10:00)	1 if after 10:00 AM and before 11:00 AM, 0 otherwise
Hour 11 Indicator (11:00)	1 if after 11:00 AM and before 12:00 PM, 0 otherwise
Hour 12 Indicator (12:00)	1 if after 12:00 PM and before 1:00 PM, 0 otherwise
Hour 13 Indicator (13:00)	1 if after 1:00 PM and before 2:00 PM, 0 otherwise
Hour 14 Indicator (14:00)	1 if after 2:00 PM and before 3:00 PM, 0 otherwise
Hour 15 Indicator (15:00)	1 if after 3:00 PM and before 4:00 PM, 0 otherwise
Hour 16 Indicator (16:00)	1 if after 4:00 PM and before 5:00 PM, 0 otherwise
Hour 17 Indicator (17:00)	1 if after 5:00 PM and before 6:00 PM, 0 otherwise
Hour 18 Indicator (18:00)	1 if after 6:00 PM and before 7:00 PM, 0 otherwise
Hour 19 Indicator (19:00)	1 if after 7:00 PM and before 8:00 PM, 0 otherwise
Hour 20 Indicator (20:00)	1 if after 8:00 PM and before 9:00 PM, 0 otherwise
Hour 21 Indicator (21:00)	1 if after 9:00 PM and before 10:00 PM, 0 otherwise
Hour 22 Indicator (22:00)	1 if after 10:00 PM and before 11:00 PM, 0 otherwise
Hour 23 Indicator (23:00)	1 if after 11:00 PM and before 12:00 PM, 0 otherwise

Table 3: Continuous independent variables used in this paper

Variable Name	Variable Description
Apparent Temperature	Numerical value representing apparent ("feels like") temperature at a given time in degrees Fahrenheit
Cloud Cover	Numerical value between 0 and 1 (inclusive) representing percentage of sky occluded by clouds
Dew Point	Numerical value representing dew point at a given time in degrees Fahrenheit
Relative Humidity	Numerical value between 0 and 1 (inclusive) representing relative humidity
Temperature	Numerical value representing temperature at a given time in degrees Fahrenheit
Wind Speed	Numerical value representing wind speed in miles per hour

researchers collaborated with the bike sharing company and developed the program in 2015. Given it is a program operated and managed internally, USF researchers had full access to the usage data, including trajectory data, of the program. With built-in GPS and the application developed by Social Bicycles, the trip data (trajectory of bikes) of each usage of the bikes is recorded in the operation management system. All trips have a unique ID. Further, each trip has a user ID, bike ID, starting timestamps, starting latitude, starting longitude, ending timestamps, ending latitude, ending longitude, trip duration (in minutes) and trip distance (in miles). Thus, the SABB program provided the perfect setting to test our proposed method. The time frame of this study was from August 28, 2015, the launch date of the program to March 30, 2017. During this time frame, a total of $189,082$ trips were recorded. However, many of these trips were noise; hence, they had to be identified and subsequently removed before any further analysis could be conducted. Trips with the following properties were removed:

- if trip duration ≤ 30 seconds, in such case, the user might be checking the bike without using it.
- if trip duration $\geq 1.5\times$ inter-quantile range of the trip duration + mean of trip duration, in such case, the user might have forgotten to lock the bike after completion of the trip.
- if trip distance $\leq .000621371$ miles or 1 meter, in such case, the bike might be damaged after short usage and the user may not able to complete his/her trip.
- if the trip either started or ended outside the USF, Tampa campus.
- if the trip is owing to a rebalancing operation.
- if the trip was conducted for testing the system.

After removing trips with the above mentioned properties, there was a total of $147,438$ trips. From this cleaned trip data, first daily and hourly dropoffs and pickups were extracted, followed by

Table 4: Quantiles of continuous variables

Continuous Variables	**Quantile**				
	Zeroth	First	Second	Third	Fourth
Apparent Temperature	28.11	67.25	75.09	82.495	107.23
Cloud Cover	0.0	0.03	0.1	0.22	1.0
Dew Point	16.55	58.16	66.0	73.08	82.14
Relative Humidity	0.16	0.62	0.79	0.89	1.0
Temperature	35.61	67.25	75.09	80.37	94.99
Wind Speed	0.0	3.87	5.66	7.82	26.55

hourly imbalance. In the case of dropoffs and pickups, their corresponding time was the starting timestamps and the ending timestamps of that particular trip respectively. From the respective timestamps, the nominal temporal variables *Season, Month, Day and Hour* were computed using date and time functions in the Julia standard library [6] and to check whether it was a holiday, the `BusinessDays.jl` package [3] was used. Once the nominal temporal variables were created, they were converted into binary (or dummy) variables, to prevent erroneous statistical estimation.

Daily and hourly weather data for the USF, Tampa campus from August 28, 2015 to March 30, 2017 were obtained using the dark sky api [4], which offers historical weather data for both daily and hourly time-frames. [4] is backed by a wide range of data sources, which are detailed in [5]. Daily and hourly weather data were then joined with the daily and hourly dropoffs and pickups as well as hourly imbalance data to obtain the final data that was used for the statistical analysis in this paper.

Decomposing continuous independent variables

Each continuous variable was decomposed into four binary variables, each of which represents a quantile range. For example, if we have a continuous variable ContVar whose quantiles are Q_1, Q_2, Q_3, Q_4, Q_5, we create four binary variables ContVar 1, ..., ContVar 4, such that ContVar 1 = 1 if $Q_1 \leq$ ContVar $< Q_2$, 0 otherwise. Table 4 describes the quantiles of the six continuous variables. Thus when $36.51^oF \leq$ Temperature $< 67.25^oF$, Temperature 1 = 1 and Temperature 2 = Temperature 3 = Temperature 4 = 0.

This operation has four major advantages. First, binary variables are easier to interpret. Second, a continuous variable by itself may not be statistically significant but one of its corresponding binary variables may be. This is in fact true in the case of the SABB dataset and is demonstrated in Section 5. Third, adding such binary variables in (quasi-) Poisson and linear regression models may improve their out-of-sample performance. This is again true in case of the SABB dataset and is demonstrated in Section 4. Finally, it is difficult to derive interactions between independent variables if one or more are continuous. So, adding binary variables corresponding to continuous variables make interactions involving continuous variables indirectly possible.

Interactions between binary independent variables

Now that we have made sure that there are binary variables corresponding to each continuous variable, we can proceed to derive interaction among binary variables. In this paper, we refer to the product of any two or any three independent binary variables, as second order and third order interactions respectively. If $\text{BinVar } 1, \text{BinVar } 2, \text{BinVar } 3$ are three independent primary binary variables, $\text{BinVar } 1 \times \text{BinVar } 2, \text{BinVar } 2 \times \text{BinVar } 3, \text{BinVar } 3 \times \text{BinVar } 1$ and $\text{BinVar } 1 \times \text{BinVar } 2 \times \text{BinVar } 3$ are second and third order interactions respectively of the three independent binary variables. Further, by definition all second and third order interactions are also binary variables.

It is important to note that, some of the above mentioned second and third order interactions will have zero variance. Such interactions should not be considered. Any interactions between binary variables for the same original variable will have zero variance, i.e, the product of any two season indicator variable will have zero variance. The same holds true for binary/indicator variables corresponding to continuous variables. Further, to prevent creation of unnecessary interactions, interactions between season and month, weekends and holiday are not considered. To ease in the variable selection procedure, certain interactions whose variance is below a predetermined threshold may also be removed. However, we do not employ any such procedure in this paper.

It is also not very clear *a priori* up to what order of interactions should be considered to achieve a desirable performance. One way of determining the highest order of interactions to be considered is via discussions and inputs from the operator, the primary user of such an analysis. Another approach is by comparing the out of sample testing errors of models with different orders of interactions used for training them. The order after which the testing error starts increasing significantly is an indication of overfitting and should be chosen as the best order of interactions.

Variable sets used in this paper

In this paper, *Var Set* refers to the set of independent variables used for training a statistical model. Four such sets are considered. The first and second sets consist of only primary (binary and continuous) variables and primary variables with decomposed binary variables of the primary continuous variables respectively. The third and fourth sets consist of all variables in the second set with all feasible second order interactions and all variables in the second set with all feasible second and third order interactions respectively.

Baseline models

To study how pickups or dropoffs vary with time and or are affected by external events such as holidays or weather conditions, negative binomial regression is commonly used in the literature (Gebhart and Noland [21]). Negative binomial regression is more appropriate than Poisson regression for the SABB dataset, as the variance of both daily and hourly dropoffs and pickups is significantly larger than their respective means. Negative binomial regression, like Poisson regression, can also be modeled as a zero-inflated or a zero-truncated model. However, in this paper no such modification is required, as we are only interested in the process that generates non-zero count variables (pickups or dropoffs). To study how hourly imbalance varies with time and or is affected by external events such as holidays or weather conditions, linear regression is used. This is because, unlike dropoffs and pickups, imbalance can also assume a negative value.

Unlike linear regression, it is difficult to interpret the coefficients of the independent variables in a negative binomial regression model directly. For this purpose, two other parameters are commonly estimated for the independent variables to determine their effects on the dependent variable. They are known as elastic and marginal effects. Elasticity of an independent variable provides an estimate of the effect of a 1% change in the independent variable on the expected frequency of the dependent variable. They provide a measure of evaluating the relative impact of each independent variable in the model. However in this paper we focus on using marginal effects rather than elastic effects owing to the ease of interpretation of marginal effects over elastic effects. Marginal effects can be more easily interpreted than elastic effects, particularly for binary variables, which are extensively present in the models used in this paper. Unlike elastic effects, marginal effects measure the effect of one unit change in the independent variable on the dependent variable. For more details on negative binomial regression models, refer to Washington et al. [29]. We use the `pscl` [2] and `mfx` [1] packages in R to estimate all the negative binomial regression models and their respective average marginal effects respectively.

It is interesting to note that, when *Var Set 3* and *4* are used, the number of independent variables outnumbers the number of observations. In such a scenario, estimating the coefficients of a negative binomial regression using maximum likelihood estimation or a linear regression using least squares cannot be used. To deal with such scenarios, we propose two stage models. In the first stage, at most n statistically significant variables are selected from the set of independent variables using a variable selection method. Once a set of variables less than the number of observations has been selected, these selected variables are used to estimate either a negative binomial or a linear regression model.

Regularization

In this section we describe two regularization strategies used in this paper:

1. Least Absolute Shrinkage and Selection Operator (LASSO) Tibshirani [27]

2. ElasticNet Zou and Hastie [31]

LASSO was introduced in Tibshirani [27]. LASSO performs both shrinkage and variable selection over a set of variables to improve the prediction accuracy and interpretability of the model. Despite having some attractive properties and features, LASSO has some disadvantages that may end up being problematic for this study. For example, if there are correlated variables, LASSO will arbitrarily select only one variable from a group of correlated variables.

ElasticNet, in certain instances, may be a better choice for regularization than LASSO, because of its above mentioned limitations. ElasticNet incorporates both L1 and L2 regularization which makes the coefficients of correlated variables shrink towards each other, while retaining the feature selection property of LASSO. This often results in selection of subsets of correlated variables. This property of ElasticNet makes it a competitive choice for variable selection along with LASSO. For more details on LASSO, ElasticNet and other regularization strategies refer to James et al. [22] and Friedman et al. [19].

We use the `glmnet` [20] package in R to compute the regularization paths for both LASSO and ElasticNet for all models in this paper. The `glmnet` package has no implementation of LASSO and ElasticNet corresponding to negative binomial distribution, so we use the implementation corresponding to Poisson distribution for daily and hourly dropoffs and pickups. This does not affect the variable selection procedure, as over-dispersion does not affect the estimates for the conditional mean. This is because, the estimating equations for the coefficients of the conditional mean are equivalent for both Poisson and negative binomial regression models. Therefore the point estimates are identical for both Poisson and negative binomial regression models when using either LASSO or Elastic Net.

Two primary parameters α and λ in glmnet need to be tuned. When $\alpha = 1$, glmnet only uses L1 regularization (LASSO) and when $0 < \alpha < 1$, glmnet uses a combination of L1 and L2 regularization (ElasticNet). Thus we vary α from 0.1 to 1.0 with a step size of 0.1. The parameter λ for both LASSO and ElasticNet is selected using 5-fold cross validation. All other parameters in glmnet are set to its default values.

Models used in this paper

Three distinct models *Model 1*, *Model 2* and *Model 3* are used in this paper. In case of daily and hourly dropoffs and pickups, *Model 1* refers to the commonly used negative binomial regression model in the literature. In case of hourly imbalance, *Model 1* refers to the linear regression model. *Model 1* is valid only for *Var Sets 1* and *2* as for *Var Sets 3* and *4* the number of independent variables is greater than the number of observations. The other two models *Model 2* and *Model 3* used in this paper are two stage models. In the first stage, a regularization strategy is used to select at most n statistically important variables from the respective variable set. This is then followed by either negative binomial regression for dropoffs and pickups or linear regression for imbalance on the set of selected variables. The first stage in *Model 2* and *Model 3* is using LASSO ($\alpha = 1$) and ElasticNet ($0 < \alpha < 1$) as the respective regularization strategy.

Model selection

Various metrics can be used to measure the quality of a negative binomial regression model. Two commonly used metrics are ρ^2 and out of sample testing error. ρ^2 statistic, also sometimes referred to as the McFadden ρ^2 is $1 - \dfrac{LL(\beta)}{LL(0)}$ where $LL(\beta)$ is the log-likelihood at convergence and $LL(0)$ is the initial log-likelihood. The ρ^2 statistic for a negative binomial regression model is always between zero and one. The closer it is to one, the better the model is. Similarly, the two most commonly used metrics for selecting linear regression models are Adjusted R^2 and out of sample testing error. The Adjusted R^2 statistic for a linear regression model is always between zero and one. The closer it is to one the better the model is.

Although ρ^2 and Adjusted R^2 statistics for negative binomial and linear regression are commonly used and provide some valuable information about the quality of a model, they fail to ascertain how well the model generalizes out of the training set. In other words, these metrics are unable to detect overfitting as they measure the quality of the model on the training set. Thus, the other measure, i.e., the root mean square error (RMSE) of the models on the hold out / testing set will be used for selecting the final models.

The dataset used in this paper, is split into two sets, the training and the testing set. The training set is used for estimating the models and comprises of trips from August 28, 2015 to February 28, 2017. The testing set is used for selecting the models. It measures how well the models generalizes out of the training set. It comprises of trips from March 1, 2017 to March 30, 2016.

Experimental Results

This section summarizes the experimental results of the proposed methods on the SABB FFBS dataset. Tables 5 and 6 summarizes the training and testing error measures for all statistical models of dropoffs and pickups and of imbalance respectively. Tables 7 and 8 reports the total number of variables and the number of variables selected corresponding to each model of dropoffs and pickups and of imbalance respectively. In Tables 7 and 8, *Vars Sel* and *SS Vars* refers to number of variables selected and the number of statistically significant variables (with 90% confidence intervals) among the variables selected for the corresponding model respectively.

Models in this paper were selected based on their testing errors, because they are a better indicator of how a model performs out of the training set, i.e., how well it generalizes out of the training set. Needless to say, the lower the testing error, the better the model is. However, if two models have similar testing errors, their training error measures can be used for breaking the tie. Unlike the testing error measure, the higher the ρ^2 or Adjusted R^2 of a model the better it is. The best models for each category are summarized in Table 9 based on the results from Tables 5 and 6.

From Tables 5 and 6, it is evident that *Var Set 2* always performs better than *Var Set 1* for all models on the SABB dataset. This indicates that it is advantageous to use *Var Set 2* instead of *Var Set 1* for training a model with no interactions on the SABB dataset, as opposed to the current trend in the literature. We also observe that, *Model 3* outperforms *Model 2* when the dependent variable is a count variable (dropoffs and pickups) except for daily dropoffs. However, the reverse is true when the dependent variable is a real number (imbalance). This indicates that 1) it is always advantageous to use either *Model 2* or *Model 3* instead of *Model 1* for training a model on the SABB FFBS dataset and 2) for training models related to dropoffs and pickups, *Model 3* is the recommended option whereas for training models related to imbalance, *Model 2* is

Table 5: Summary of training and testing error measures for all models of dropoffs and pickups

Variable	Time-frame	Var Set	Model Used					
			Model 1		Model 2		Model 3	
			ρ^2	RMSE	ρ^2	RMSE	ρ^2	RMSE
Dropoffs	Daily	1	0.0438	256.9260	0.0362	189.5043	0.0366	188.2041
		2	0.0470	253.3899	0.0439	152.6411	0.0378	**150.3104**
		3	-		0.0702	224.1921	0.0617	231.6895
		4			0.0854	**148.4511**	0.0873	186.2352
Pickups		1	0.0437	256.8616	0.0437	256.8616	0.0365	184.3904
		2	0.0470	253.3143	0.0378	150.3913	0.0378	**150.3913**
		3	-		0.0620	231.1562	0.0661	252.7010
		4			0.0955	190.7476	0.0903	**144.1414**
Dropoffs	Hourly	1	0.1161	11.9317	0.1161	11.9317	0.1161	11.9317
		2	0.1179	11.2325	0.1179	11.2325	0.1179	**11.2325**
		3	-		0.1668	18.7437	0.1668	18.7437
		4			0.1945	15.1279	0.1915	**14.5176**
Pickups		1	0.1159	11.9516	0.1159	11.9516	0.1159	11.9516
		2	0.1178	11.2552	0.1178	11.2552	0.1178	**11.2552**
		3	-		0.1667	17.2632	0.1667	17.2632
		4			0.1982	14.5979	0.1940	**14.0161**

Table 6: Summary of training and testing error measures for all models of hourly imbalance

Variable	Time-frame	Var Set	Model Used					
			Model 1		Model 2		Model 3	
			Adjusted R^2	RMSE	Adjusted R^2	RMSE	Adjusted R^2	RMSE
Imbalance	Hourly	1	0.0422	0.6503	0.0442	0.6484	0.0441	0.6487
		2	0.0420	0.6495	**0.0444**	**0.6483**	0.0444	0.6484
		3	-		**0.1250**	**0.7262**	0.1259	0.7448
		4			0.1857	0.7326	0.1857	0.7326

Table 7: Summary of variable selection for all models of dropoffs and pickups

Variable	Time-frame	Var Set	Total Vars	Model Used					
				Model 1		Model 2		Model 3	
				Vars Sel	SS Vars	Vars Sel	SS Vars	Vars Sel	SS Vars
Dropoffs	Daily	1	27	27	19	12	6	16	6
		2	44	44	19	33	16	**18**	**11**
		3	928	-		100	36	70	23
		4	8160			**127**	**45**	132	44
Pickups		1	27	27	19	27	19	14	7
		2	44	44	19	18	11	**18**	**11**
		3	928	-		75	23	89	26
		4	8160			160	51	**149**	**45**
Dropoffs	Hourly	1	50	50	47	50	47	50	47
		2	66	66	57	66	57	**66**	**57**
		3	2146	-		922	378	922	378
		4	31734			1348	617	**1271**	**578**
Pickups		1	50	50	46	50	46	50	46
		2	66	66	57	66	56	**66**	**56**
		3	2146	-		906	371	906	371
		4	31734			1486	695	**1350**	**617**

Table 8: Summary of variable selection for all models of imbalance

Variable	**Time-frame**	**Var Set**	**Total Vars**	**Model Used**					
				Model 1		Model 2		Model 3	
				Vars Sel	**SS Vars**	**Vars Sel**	**SS Vars**	**Vars Sel**	**SS Vars**
Imbalance	**Hourly**	**1**	50	50	12	18	14	19	14
		2	66	66	19	**24**	**16**	23	15
		3	2146	-		**184**	**131**	201	133
		4	31734			170	137	170	136

Table 9: Selected models

Variable	**Time-frame**	**Selected Model**	
		No Interactions	**With Interactions**
Dropoffs	**Daily**	*Model 3* with *Var Set 2*	*Model 2* with *Var Set 4*
Pickups		*Model 3* with *Var Set 2*	*Model 3* with *Var Set 4*
Dropoffs	**Hourly**	*Model 3* with *Var Set 2*	*Model 3* with *Var Set 4*
Pickups		*Model 3* with *Var Set 2*	*Model 3* with *Var Set 4*
Imbalance		*Model 2* with *Var Set 2*	*Model 2* with *Var Set 3*

the recommended option.

Another interesting observation is that, the sparsest model is always performing the best. By the sparsest model, we refer to the model whose *Vars Sel* is the lowest. This in a way is an indication that the simpler the model is, the better it tends to perform. Hence, we can conclude that two stage models proposed in this paper, generates models that are not only simple/sparse (models with fewer number of variables) but also closer to the ground truth (as their testing errors are lower) than the baseline *Model 1* with *Var Set 1*, commonly used in the literature. It is interesting to note that, when interactions are added to the model, it sometimes performs better than models with no interactions and sometimes does not. However, it is almost always true that the quality of the model improves when the order of the interactions is increased, except for hourly imbalance. Although, we limit ourselves to third order interactions in this paper, this indicates that increasing the order of the interactions from third to fourth or even fifth may improve the quality of the model, but it will come at a higher cost of computational complexity and difficultly in interpreting the resulting model.

Adding interactions does not always improve the testing error of a model (it always improve the training error). For example: from Table 5, it is evident that for daily time-frame, the best models with interactions outperform the best models without interactions, however the same cannot be said for hourly time-frames. This leads to some interesting insights. For daily time-frame, *Model 2* and *Model 3* with *Var Set 4* for dropoffs and for pickups respectively, have some third order interactions (mentioned in Table 10) which by themselves are not statistically significant in *Model 3* with *Var Set 2* for both dropoffs and pickups. This is a clear indication that the best models with interactions are able to capture information, which were missed by the corresponding best models with no interactions. This characteristic of the best models with interactions being able to capture information that the best models without interactions cannot becomes more evident in Section 5.3.

Table 10: Variables that become significant when combined together

Independent Variable	Time-frame	Dependent Variables		
		Variable 1	**Variable 2**	**Variable 3**
Dropoffs	**Daily**	Spring	Wind Speed 2	Cloud Cover 3
		September	Tuesday	Cloud Cover 3
		February	Tuesday	Relative Humidity 4
		Spring	Temperature 2	Cloud Cover 2
		Monday	Cloud Cover 2	Relative Humidity 2
		September	Wind Speed 3	Cloud Cover 2
		September	Temperature 4	Wind Speed 2
		Tuesday	Cloud Cover 2	Relative Humidity 4
		Tuesday	Temperature 1	Wind Speed 3
		February	Monday	Cloud Cover 4
Pickups		September	Tuesday	Cloud Cover 3
		February	Wind Speed 1	Cloud Cover 1
		November	Wind Speed 1	Cloud Cover 2
		February	Tuesday	Relative Humidity 4
		October	Dew Point 2	Cloud Cover 4
		September	Temperature 4	Wind Speed 2
		September	Dew Point 3	Relative Humidity 4
		February	Monday	Cloud Cover 4
		Tuesday	Cloud Cover 2	Relative Humidity 4
		Apparent Temperature 3	Dew Point 3	Cloud Cover 1

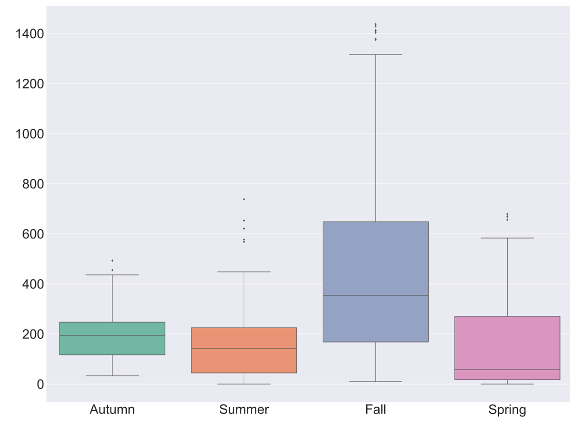
Thus, it important that instead of choosing a model with or without interactions over another, both models are used in conjunction to complement each other weaknesses with their strengths.

Discussion

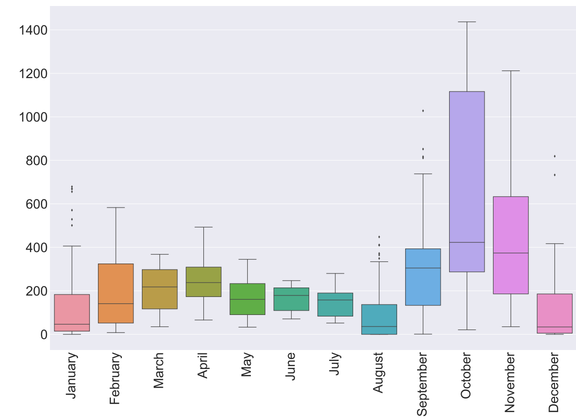
In this section, we demonstrate how to interpret and draw inferences from visualization of historical data, best models with no interactions, best models with interactions and by combining all these methods. Then, we demonstrate how to provide appropriate recommendations to the operator, based on these respective inferences. In this paper, we use only pickups and imbalance for drawing inferences and providing recommendations. The reason for this is two-fold: 1) to prevent repetition and 2) in the case of free-floating systems, dropoffs have very little effect on the demand of system as they have no explicit (capacity) restriction, unlike in the case of station-based systems. Further, pickups for both free-floating and station-based systems is a far better indicator of the approximate demand of the system. In case of station-based systems, dropoffs may also be considered in conjunction to pickups.

Data Visualization

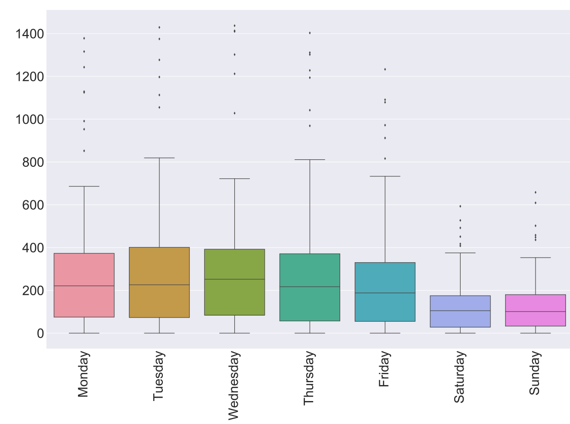
Figures 1a through 1d visualize how daily pickups vary with season, month, day and holiday respectively, in the SABB dataset. Figures 1e and 1f visualize how hourly pickups and imbalance vary with hours in a day respectively, in the SABB dataset. From Figures 1a and 1b, we can infer



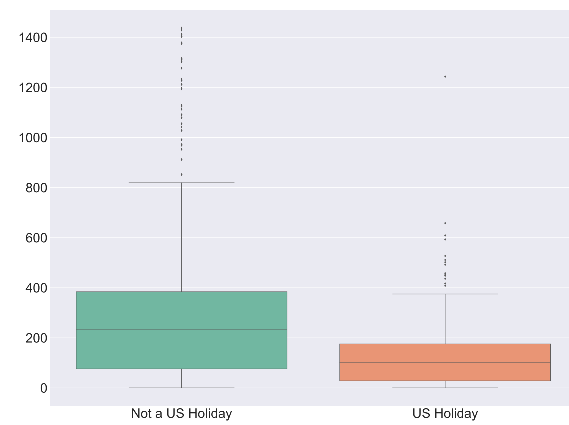
(a) Variation of Daily Pickups with Season



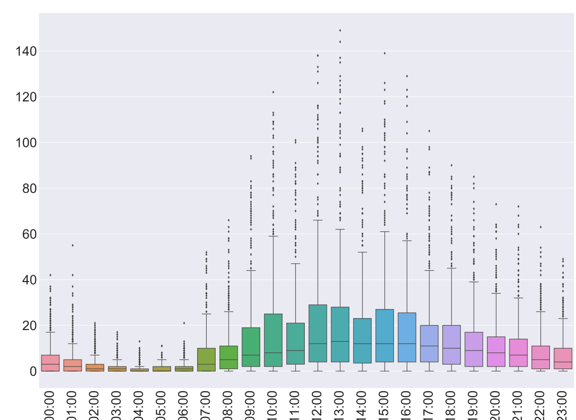
(b) Variation of Daily Pickups with Month



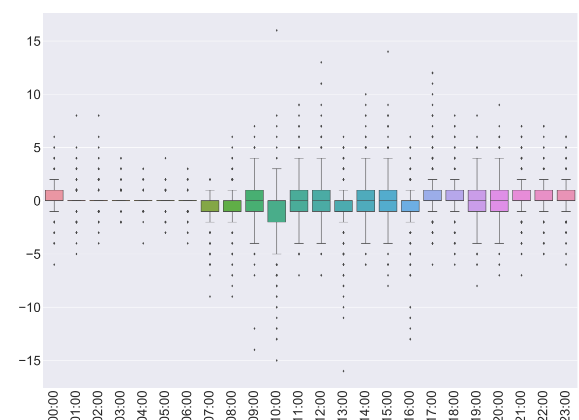
(c) Variation of Daily Pickups with Day



(d) Variation of Daily Pickups with Holiday



(e) Variation of Hourly Pickups with Hour



(f) Variation of Hourly Imbalance with Hour that there is significant variation in pickups owing to both season and month. The two primary causes for this phenomenon, are the correlation of both season and month with the timing of semesters at USF and weather conditions. Most trips are reported in the Fall semester, when the weather is pleasant. There is a dip in usage for both the Spring and Summer semesters because the weather in the beginning of both of these semesters is a bit more severe compared to that in the fall semester. Further, fewer students are present on campus during the Summer semester. From Figures 1c and 1d, we can conclude that pickups are higher on weekdays than on weekends or holidays. This is owing to more activity (inter class or dorm to class or class to dorm trips) on campus on weekdays than on weekends. Pickups are maximum on Tuesday, followed by Wednesday, Monday, Thursday and Friday. This is because, most USF classes are held on Tuesday, followed by Wednesday, Monday, Thursday and Friday. From Figure 1e, we can conclude that pickups start increasing at 7:00 AM (when classes start), and peak around 1:00 PM. From Figure 1f, we can conclude that there is negative imbalance in the system from 7:00 AM to 9:00 AM, 10:00 AM to 11:00 AM, 1:00 PM to 2:00 PM and 4:00 PM to 5:00 PM. This phenomenon is because of class timings and extracurricular activity patterns of students and staff at USF. Based on Figures 1a through 1f, we recommend to the operator of the SABB FFBS that, the best time-frame for static rebalancing or on-site maintenance is 1:00 AM to 7:00 AM, because the pickups on average are almost close to zero during this time period and the appropriate time-frames for dynamic rebalancing are 9:00 AM to 10:00 AM, 11:00 AM to 1:00 PM and 2:00 PM to 4:00 PM.

Models with No Interactions

Figures 2 and 3, visualize the average marginal effects of statistically significant variables for the best models with no interaction for daily and hourly pickups respectively. From Figures 2 and 3, we can conclude that fall season (and its corresponding months) has a significant positive impact on both daily and hourly pickups. On the contrary, for both Spring and Summer seasons and for their corresponding months, there is a sudden dip for both daily and hourly pickups. From figure 3, it is clear that 11:00 AM to 12:00 PM is the peak time frame, which is a bit different than that obtained from data visualization. Further, the time frames 7:00 AM to 9:00 PM and 11:00 PM to 6:00 AM have a positive and a negative impact on hourly pickups respectively. It is not a surprise that both daily and hourly pickups decrease on holidays. It is interesting to note that, even though *dew point* and *wind speed* by themselves are not statistically significant, when the dew point is $16.55 - 66.0^oF$ and when wind speed is between $5.66 - 26.55$ mph they not only become statistically significant but also negatively impact hourly pickups. Further, hourly pickups decrease as the sky becomes more clouded, because it is less likely for users to commute using bikes when there is a high possibility of raining. Another interesting phenomenon occurs in the case of relative humidity. Relative humidity by itself negatively impacts hourly pickups, as it is a measure of extreme conditions. However, when relative humidity is either $0.16 - 0.62$ or $0.79 - 0.89$, pickups increase significantly. It is important to note that, we are able to identify these intervals for *dew point*, *wind speed* and *relative humidity* because of our proposed variable decomposition strategy.

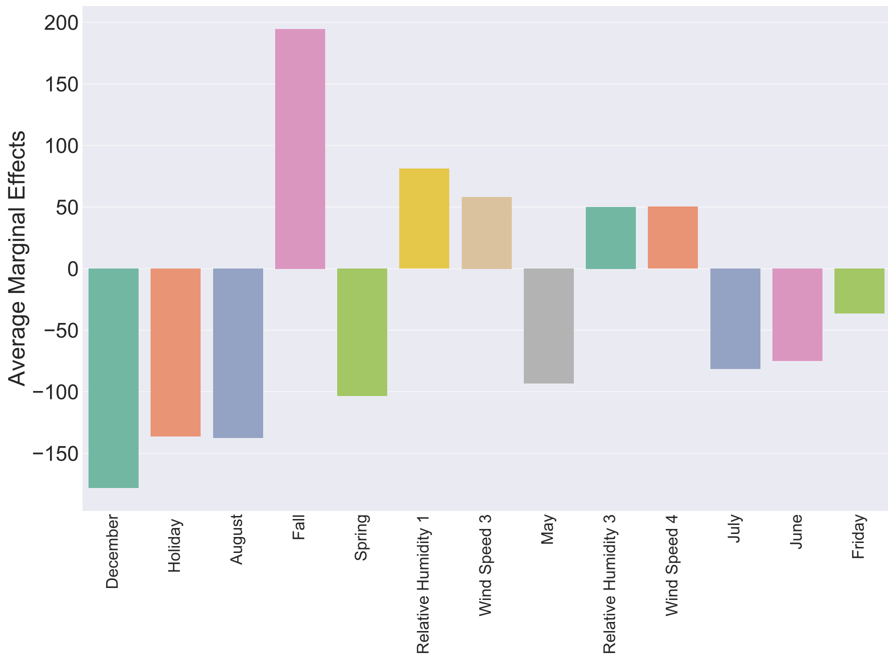


Figure 2: Average marginal effects of statistically significant variables for the best model with no interactions for daily pickups

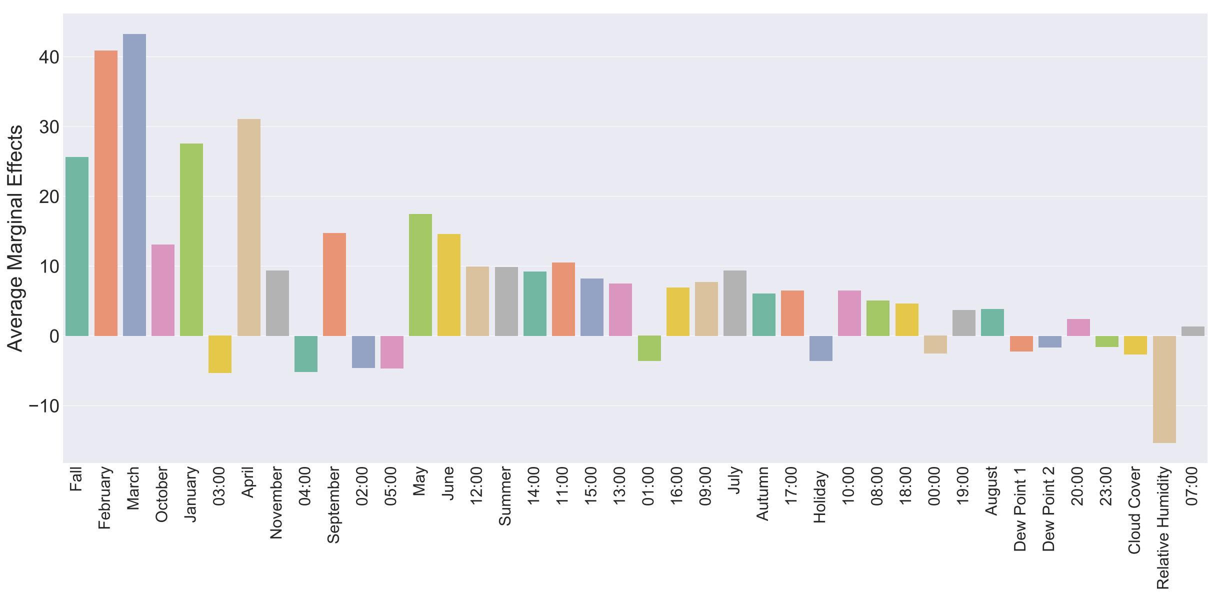


Figure 3: Average marginal effects of statistically significant variables for the best model with no interactions for hourly pickups

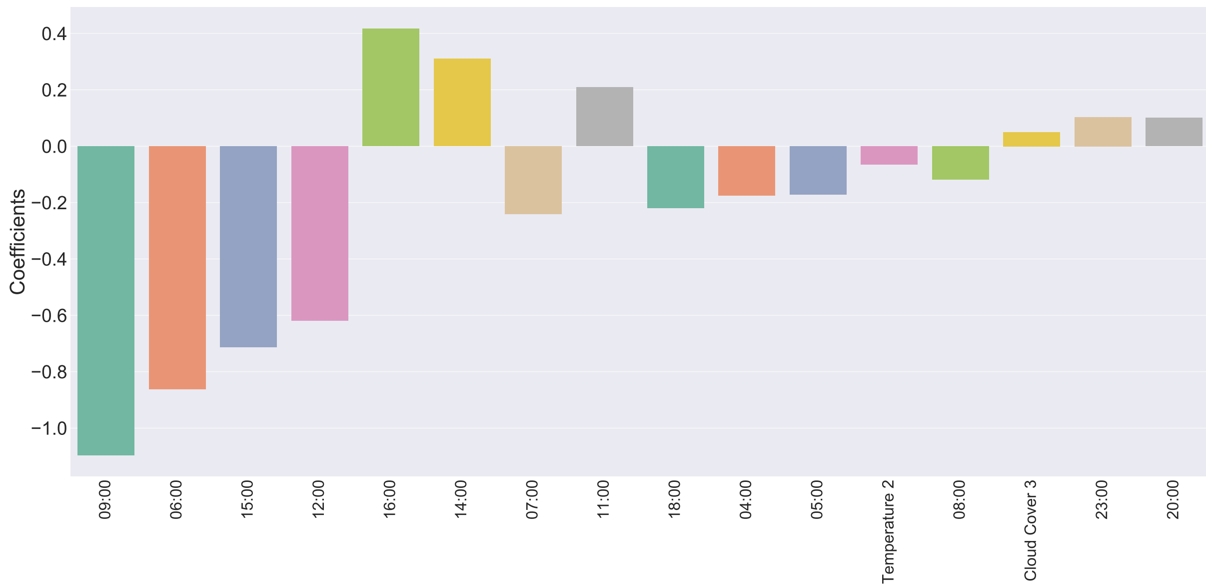


Figure 4: Coefficients of statistically significant variables for the best model with no interactions for hourly imbalance

Figure 4, visualize the coefficients of statistically significant variables for the best model with no interaction, for hourly imbalance. Figure 4 gives a clear indication of the time-frames of interest when imbalance is negative, i.e., 6:00 AM to 10:00 AM, 12:00 PM to 1:00 PM and 3:00 PM to 4:00 PM. Thus, based on Figures 2, 3 and 4, we can provide the following three recommendations. First, (operator-based) static rebalancing and on-site maintenance operations can be conducted between 11:00 PM - 6:00 AM on a desired day. Second, dynamic rebalancing (both operator-based and user-based) if required should be held between the hours of 7:00 AM to 8:00 AM, 10:00 AM to 12:00 PM and 1:00 PM to 3:00 PM. Finally, we recommend the operator to use a user-based dynamic rebalancing / user incentives schemes in the Spring, in May, June, July, August and December, on Fridays and on holidays.

Models with Interactions

Figures 5 and 6, visualizes the average marginal effects of first order statistically significant variables for the best models with interactions, for both daily and hourly pickups respectively. From figures 2 and 3, we can conclude that fall season has a significant positive impact on both daily and hourly pickups. Similarly, December has a negative impact on both daily and hourly pickups. This is because many students return to their homes during this time after the semester has concluded. Thus there is a dip in the number of users. March and April as well as, October have a positive and a negative impact on pickups respectively. From figure 3, it is clear that 11:00 AM to 12:00 PM is the peak time frame, with the time frame 9:00 AM to 3:00 PM and 11:00 PM to 6:00 AM having a positive and a negative impact on hourly pickups respectively. It is not surprising that

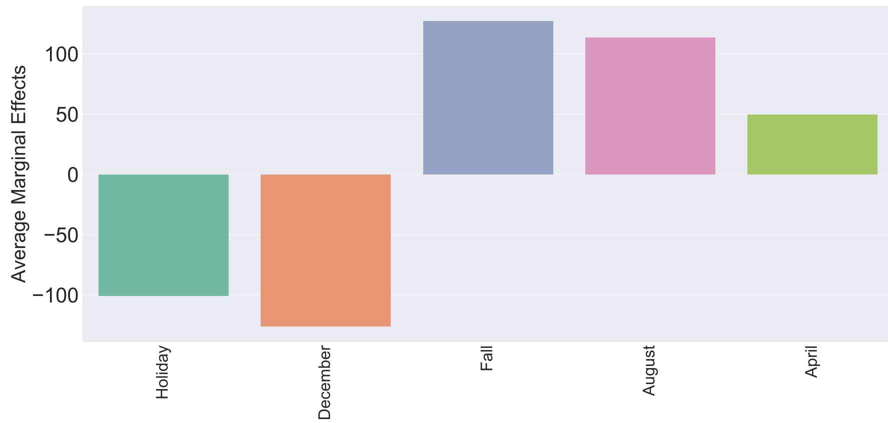


Figure 5: Average marginal effects of first order statistically significant variables for the best model with interactions for daily pickups

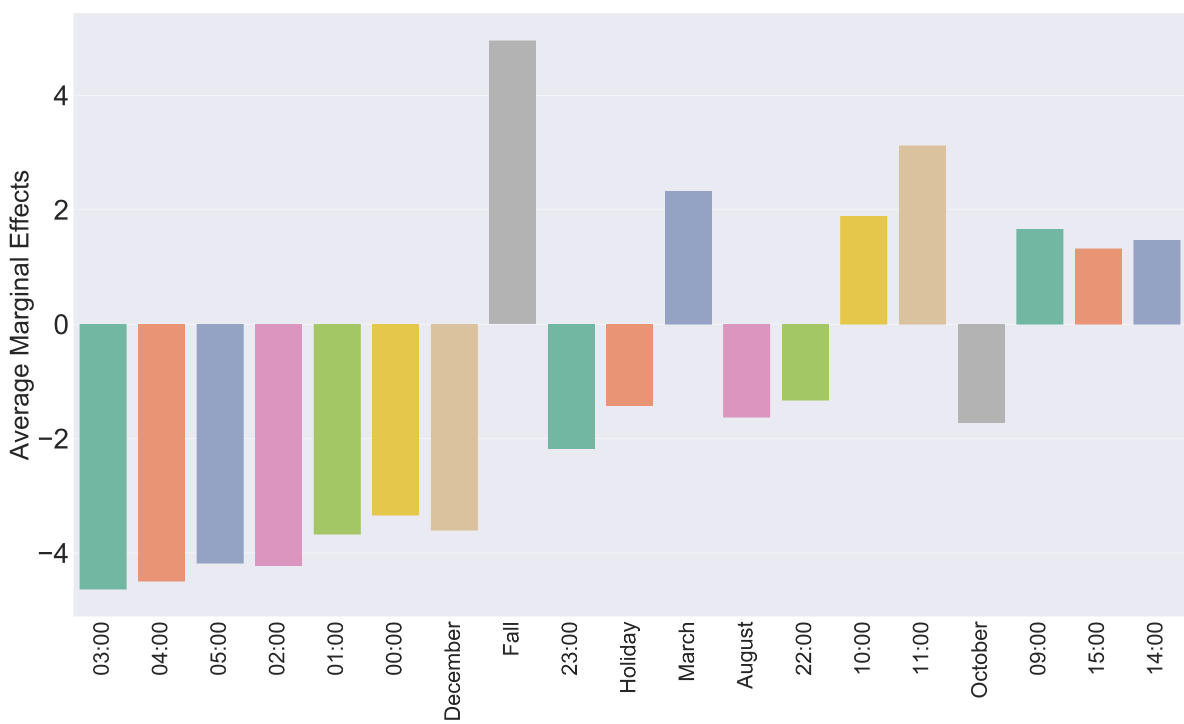


Figure 6: Average marginal effects of first order statistically significant variables for the best model with interactions for hourly pickups

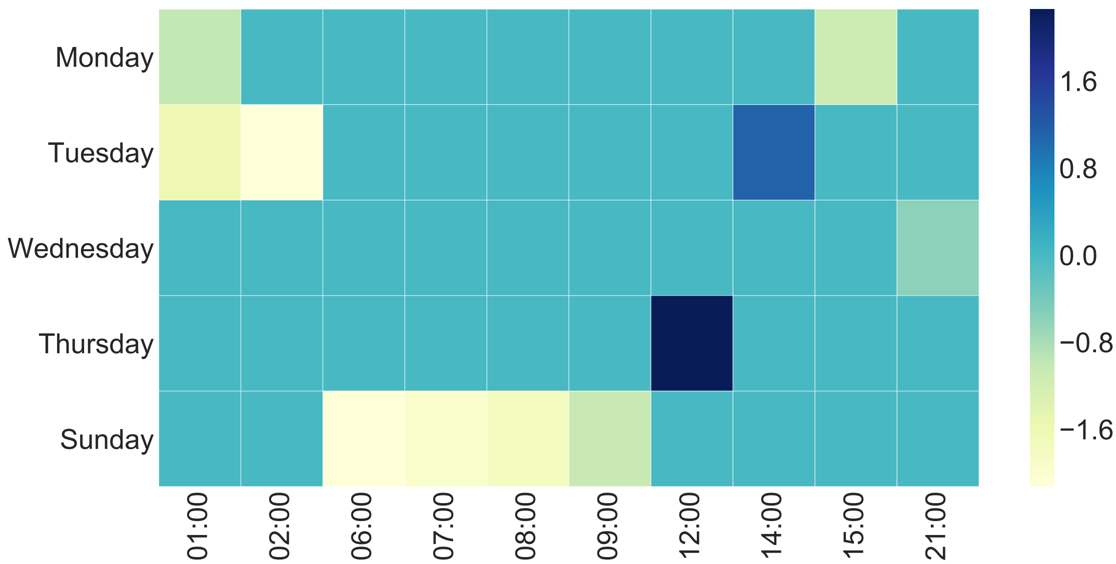


Figure 7: Average marginal effects of second order statistically significant variables between day, holiday and hour for the best model with interactions for hourly pickups

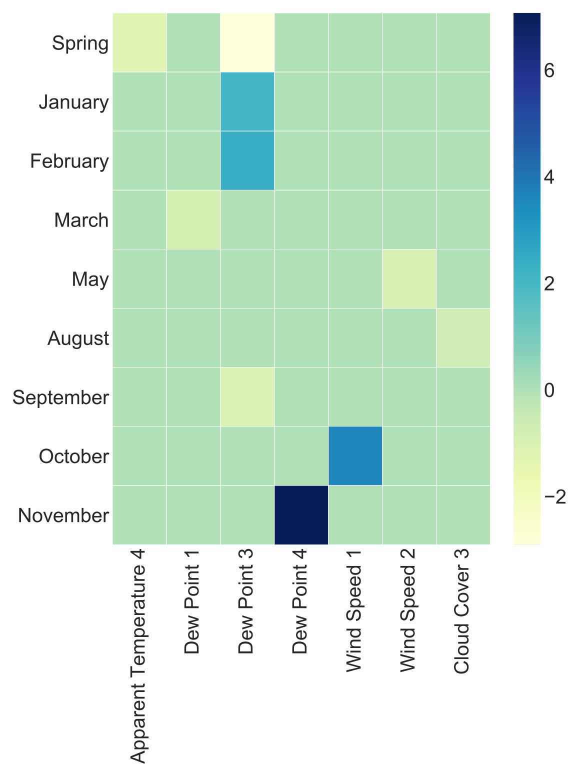
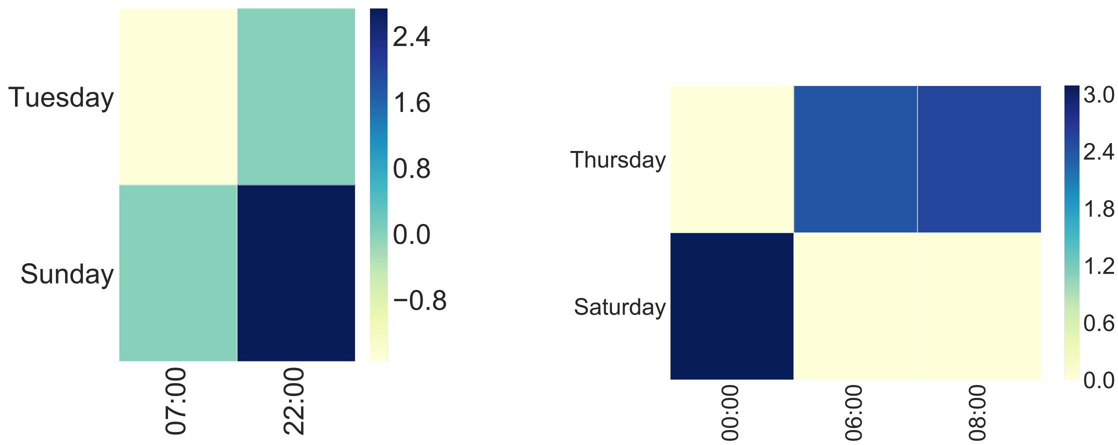


Figure 8: Average marginal effects of second order statistically significant variables between season, month and weather variables for the best model with interactions for hourly pickups



(a) September (b) October

Figure 9: Average marginal effects of third order statistically significant variables between September/October, day, holiday and hour for the best model with interactions for hourly pickups

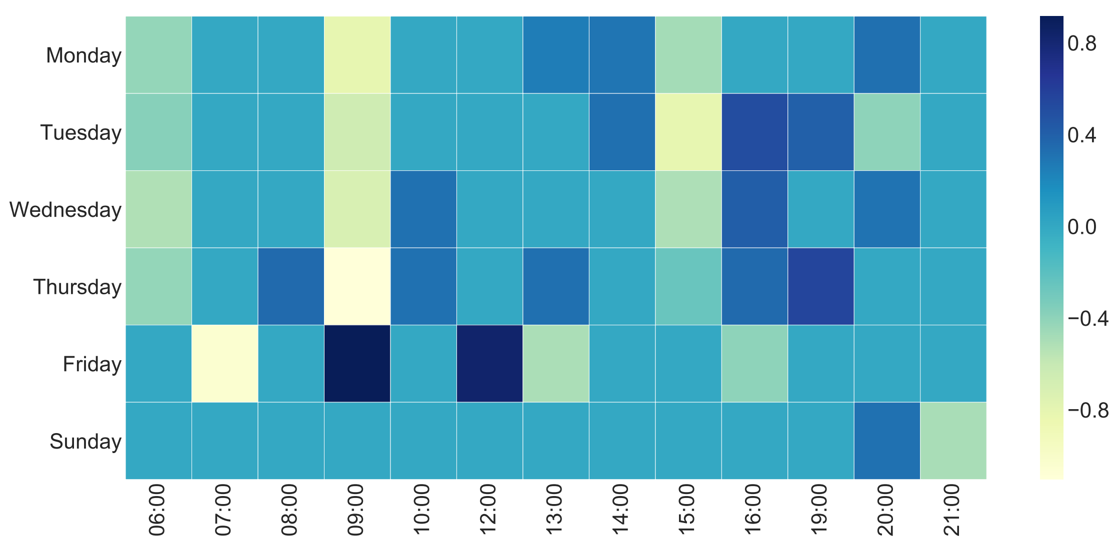


Figure 10: Coefficients of second order statistically significant variables between day, holiday and hour for the best model with interactions for hourly imbalance

both daily and hourly pickups decrease during holidays.

Figures 7 and 8, visualize the average marginal effects of second order statistically significant variables between day, holiday and hour variables and between season, month and weather variables for the best model with interactions for hourly pickups respectively. From figure 7, we can make some interesting conclusions. First, there is a sudden drop in pickups on Mondays from 3:00 PM to 4:00 PM. Second, there is a sudden increase in pickups on Tuesdays from 2:00 PM to 3:00 PM. Finally, on Thursdays there is a sudden increase from 12:00 PM to 1:00 PM. Perhaps be on Thursdays the peak is from 12:00 PM to 1:00 PM instead of from 11:00 AM to 12:00 PM. From figure 8, we can make some interesting conclusions. When the apparent temperature is $82.495 - 107.23^oF$ during Spring, there is a decrease in hourly pickups. When the dew point is $16.55 - 58.16^oF$ during March, there is a decrease in hourly pickups. When the dew point is $66.00 - 73.08^oF$, there is a decrease in hourly pickups during Spring and during September, whereas the hourly pickups increases during the months of January and February. When the dew point is $73.08 - 82.14^oF$ during November, there is an increase in hourly pickups. When the wind speed is $0.00 - 3.87$ mph during October, hourly pickups increase. When the wind speed is $3.87 - 5.66$ mph during May, hourly pickups decrease. When the cloud cover is $0.1 - 0.22$ during August, hourly pickups decrease.

Figures 9a and 9b, visualize average marginal effects of third order statistically significant variables between September/October, day, holiday, and hour for the best model with interactions for hourly pickups respectively. From Figure 9a, we can conclude that in September, Tuesdays have a slower start compared to other months and on Sundays, there is an increase in pickups during 10:00 PM to 11:00 PM. From figure 9b, we can conclude that in October, Thursdays have an early start at 6:00 AM instead of at 7:00 AM, and on Saturdays there is a increase in pickups during 12:00 AM to 01:00 AM. The increase in pickups from 10:00 PM to 11:00 PM on Sundays in September and from 12:00 AM to 01:00 AM on Saturdays during October, may be because of students engaging in recreational activities during weekends in the middle of the fall semester.

Figure 10 visualizes the coefficients of second order statistically significant variables between day, holiday and hour for the best model with interactions for hourly imbalance. This figure provides a lot of valuable information. First, the trend of imbalance on a Friday is quite different from that on the other weekdays. Clearly, during 6:00 AM to 7:00 AM, 9:00 AM to 10:00 AM and 3:00 PM to 4:00 PM on Monday to Thursday there is negative imbalance in the system. On Friday, the negative imbalance is during 7:00 AM to 8:00 AM, 1:00 PM to 2:00 PM and 4:00 PM to 5:00 PM. This phenomenon arises due to the difference in class schedules on Friday compared to that on the other weekdays. On Sunday, there is a negative imbalance from 9:00 PM to 10:00 PM, which may be because of students engaging in recreational activities.

Based on the above inferences, we can provide the following three recommendations. First, (operator-based) static rebalancing and on-site maintenance operations can be conducted between 11:00 PM - 6:00 AM on a desired day, except for Tuesdays in September when it may be extended until 8:00 AM. Second, dynamic rebalancing (both operator-based and user-based), if required should be held from 10:00 AM to 3:00 PM on Monday through Thursday and from 9:00 AM to 1:00 PM and 2:00 PM to 4:00 PM on Friday. Third, we recommend the operator to use static rebalancing strategies in Fall, in April and August and dynamic rebalancing strategies in December and on holidays.

All Vantage Points

In this section, we synthesize inferences and recommendations derived from three vantage points, namely data visualization of historical data, best models with and without interactions. An inference or a recommendation is strongest if it can be validated by all of the above three methods, and weakest if only one of the above three methods validates it. For example: based on data visualization and best models with and without interactions, the best time for static rebalancing or onsite maintenance is from 1:00 AM to 7:00 AM, 11:00 PM to 6:00 AM and 11:00 PM to 6:00 AM respectively. However, if all three of these recommendations are combined, it is clear that 1:00 AM to 6:00 AM is a time frame that is valid from all of these three methods. Similar approach is followed in this section for inferences and recommendations.

Based on the above guidelines, we can draw the following conclusions about the mobility patterns of the SABB FFBS:

1. Fall has a significant positive impact on pickups, whereas, both Spring and Summer have a negative impact on pickups.
2. March and April have a positive impact, and October and December have a negative impact on pickups respectively.
3. Pickups are higher on weekdays than on weekends or holidays, reaching a peak on Tuesday, followed by Wednesday, Monday, Thursday and Friday.
4. Peak hours are from 11:00 AM to 12:00 PM (except for Thursdays when the peak is 12:00 PM to 1:00 PM), with the time frames 9:00 AM to 3:00 PM and 10:00 PM to 6:00 AM having a positive and a negative impact on pickups respectively.
5. There is a sudden decrease in pickups on Mondays from 3:00 PM to 4:00 PM and a sudden increase in pickups from 10:00 PM to 11:00 PM on Sundays in September and from 12:00 AM to 01:00 AM on Saturdays during October.
6. There is a decrease in pickups in Spring when the apparent temperature is $82.495 - 107.23^oF$.
7. In October, pickups increase when wind speed is $0.00 - 3.87$ mph, however, pickups decrease when wind speed is $3.87 - 5.66$ mph in May and between $5.66 - 26.55$ mph.
8. Pickups decrease when the dew point is $16.55 - 66.0^oF$, or $66.00 - 73.08^oF$ in Spring and September, however pickups increase when the dew point is between $66.00 - 73.08^oF$ in January and February and between $73.08 - 82.14^oF$ in November.

9. Pickups decrease with increase in cloud cover.

10. Relative humidity by itself negatively impacts pickups, however, when relative humidity is either $0.16 - 0.62$ or $0.79 - 0.89$, pickups increase significantly.

Similarly, based on the above guidelines, it is clear that during 6:00 AM to 7:00 AM, 9:00 AM to 10:00 AM and 3:00 PM to 4:00 PM on Monday to Thursday there is negative imbalance in the system. On Friday, the negative imbalance is during 7:00 AM to 8:00 AM, 1:00 PM to 2:00 PM and 4:00 PM to 5:00 PM. By combining insights and recommendations from all vantage points, we can provide the following final recommendations to the operator of the SABB FFBS. The best time for static rebalancing or on-site maintenance is between 1:00 AM and 6:00 AM, except for Tuesdays in September when it may be extended until 8:00 AM. Dynamic rebalancing (both operator-based and user-based), if required should be held from 10:00 AM to 12:00 PM and 1:00 PM to 3:00 PM on Monday through Thursday and from 9:00 AM to 1:00 PM and 2:00 PM to 4:00 PM on Friday. Static rebalancing strategies be extensively used in Fall and in April. Dynamic rebalancing strategies should be used in May, June, July and December, and on holidays.

Conclusion

In this paper, we propose a method to extract operational management insights from historical trip data of a shared mobility system, to help the operator make more informed decisions. A significant amount of research has been conducted on gaining various forms and types of insights with a broad range of motivation, from the historical data of the system. However, none of these studies considered interaction between independent variables or study imbalance as a dependent variable. In this paper, we take interactions among independent variables into consideration and apply methods to remove unnecessary interactions. We also show that more insights about the mobility patterns and imbalance of the SABB program can be obtained by considering such interactions. We also propose a simple method to decompose continuous variables into binary variables which improves the base model used in the literature. Our proposed methodology gives a unique opportunity to study the system and make recommendations to the operator from various vantage points. To extend our proposed method for station-based systems, dropoffs can also be considered in conjunction to pickups.

Even though the two stage models perform better than baseline (quasi) Poisson regression models, their testing error measure is not as low as one would expect. A possible explanation for this effect is that both the two stage and the baseline models are linear models. Thus they are unable to capture possible non-linear relationships among the independent and the dependent variables. This effect is mitigated to some extent by adding up to third order interactions, as they are able to capture unobserved heterogeneity in the data. Adding fourth or even higher order interactions may improve the model, however doing so may make the model difficult to interpret. Thus, it is our belief that interactions higher than third order are unnecessary, instead nonlinear transformations and interactions may be added to determine if the performance of the models improves or not. This is a possible future research direction.

In future papers, we will address how to use information from such an analysis to compute optimal inventory levels, which can then be used by the operator as inputs to their specific rebalancing strategies. Another possible research direction can be conducting this analysis for each station in case of station based bike sharing systems or each zone in case of free floating bike sharing systems.

References

[1] (2015). mfx. `https://cran.r-project.org/web/packages/mfx/mfx.pdf`.

[2] (2015). pscl. `http://pscl.stanford.edu/`.

[3] (2017). BusinessDays.jl. `https://github.com/felipenoris/BusinessDays.jl`.

[4] (2017a). Dark Sky Api. `https://darksky.net/dev/docs/forecast`.

[5] (2017b). Dark Sky Data Sources. `https://darksky.net/dev/docs/sources`.

[6] (2017). Julia Stdlib - Dates and Time. `https://docs.julialang.org/en/stable/stdlib/dates/#stdlib-dates-1`.

[7] Alvarez-Valdes, R., Belenguer, J. M., Benavent, E., Bermudez, J. D., Muoz, F., Vercher, E., and Verdejo, F. (2016). Optimizing the level of service quality of a bike-sharing system. *Omega*, 62:163 – 175.

[8] Borgnat, P., Abry, P., Flandrin, P., Robardet, C., Rouquier, J.-B., and Fleury, E. (2011). Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(03):415–438.

[9] Caulfield, B., O'Mahony, M., Brazil, W., and Weldon, P. (2017). Examining usage patterns of a bike-sharing scheme in a medium sized city. *Transportation Research Part A: Policy and Practice*, 100:152 – 161.

[10] Cheu, R., Xu, J., Kek, A., Lim, W., and Chen, W. (2006). Forecasting shared-use vehicle trips with neural networks and support vector machines. *Transportation Research Record: Journal of the Transportation Research Board*, (1968):40–46.

[11] de Chardon, C. M. and Caruso, G. (2015). Estimating bike-share trips using station level data. *Transportation Research Part B: Methodological*, 78:260 – 279.

[12] de Chardon, C. M., Caruso, G., and Thomas, I. (2017). Bicycle sharing system success determinants. *Transportation Research Part A: Policy and Practice*, 100:202 – 214.

[13] DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4):41–56.

[14] Faghih-Imani, A., Anowar, S., Miller, E. J., and Eluru, N. (2017a). Hail a cab or ride a bike? a travel time comparison of taxi and bicycle-sharing systems in new york city. *Transportation Research Part A: Policy and Practice*, 101:11 – 21.

[15] Faghih-Imani, A. and Eluru, N. (2016). Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of new york citibike system. *Journal of Transport Geography*, 54:218 – 227.

[16] Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., and Haq, U. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (bixi) in montreal. *Journal of Transport Geography*, 41:306 – 314.

[17] Faghih-Imani, A., Hampshire, R., Marla, L., and Eluru, N. (2017b). An empirical analysis of bike sharing usage and rebalancing: Evidence from barcelona and seville. *Transportation Research Part A: Policy and Practice*, 97:177 – 191.

[18] Fishman, E., Washington, S., Haworth, N., and Watson, A. (2015). Factors influencing bike share membership: An analysis of melbourne and brisbane. *Transportation Research Part A: Policy and Practice*, 71:17 – 30.

[19] Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning*, volume 1. Springer.

[20] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

[21] Gebhart, K. and Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in washington, dc. *Transportation*, 41(6):1205–1225.

[22] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

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

[24] OBrien, O., Cheshire, J., and Batty, M. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, 34:262 – 273.

[25] Pal, A. and Zhang, Y. (2017). Free-floating bike sharing: Solving real-life large-scale static rebalancing problems. *Transportation Research Part C: Emerging Technologies*, 80:92 – 116.

[26] Regue, R. and Recker, W. (2014). Proactive vehicle routing with inferred demand to solve the bikesharing rebalancing problem. *Transportation Research Part E: Logistics and Transportation Review*, 72:192 – 209.

[27] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[28] Wagner, S., Brandt, T., and Neumann, D. (2016). In free float: Developing business analytics support for carsharing providers. *Omega*, 59:4 – 14. Business Analytics.

[29] Washington, S. P., Karlaftis, M. G., and Mannering, F. (2010). *Statistical and econometric methods for transportation data analysis*. CRC press.

[30] Zhang, Y., Thomas, T., Brussel, M. J. G., and van Maarseveen, M. F. A. M. (2016). Expanding bicycle-sharing systems: Lessons learnt from an analysis of usage. *PLOS ONE*, 11(12):1–25.

[31] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.