Research on Antecedents and Consequences of Factors Affecting the Bike Sharing System

--- Lessons From Capital Bike Share Program in Washington, D.C.

Chen Jing
Department of Mathematic
University of Iowa
Iowa city, USA
chen-jing@uiowa.edu

Zuoyuan Zhao
Department of Computer Science
University of Iowa
Iowa city, USA
zuoyan-zhao@uiowa.edu

Abstract—Bike sharing has drawn significant attention in academia and business. People can use bike sharing system to rent a bike at one of many rental stations traveled around the city for a short journey and return it at any station in the city. However, there have been problems raising with the application of the system. This research focuses on what factors affect the bike sharing system operation. Specifically, the research model is combined of Multiple linear regression, Poisson and Topology method. The multiple linear regression is used to analyze the qualitative data, poisson and topology method for quantitative data. Additionally, the antecedents and consequences of each aspect are examined by studying the case in 2011 Washington, D.C., the influence for each individual factor is also measured. By running the method into software R, three facets of environment are concluded as important factors for the bike sharing system. The results indicate that combination of sensory temperature, season and weather impacts the bike sharing system. At the same time, the suitable schedule for stations to reposition or repair the sharing bike can be concluded. The methodology also offers specific insights to city managers for operating the bike sharing systems.

Keywords-bike sharing system; reposition; multiple linear regression; partial regression plot; poisson; topology.

I. INTRODUCTION

Bicycle sharing systems or bike share scheme began in Europe in 1965. There are systems that provide short term bike service to individuals. The idea is to install bike stations at various points in the city, from which registered users can easily loan a bike by removing it from a specialized rack. After the ride, the users may return the bike at any arbitrary station (provided that there is a free rack) [1]. It is also becoming popular as a sustainable means of transportation in the urban environment [2]. As of June 2014, public bike sharing programs existed on a few continents, including over hundreds of cities, operating over hundreds of thousands of bicycles at over tens of thousands of stations [3]. Bike sharing systems seem to become a new fashion all around the world. In addition, by providing an incentive for doing sports they are a significant contribution to improving public health [4]. However, there are also problems raise with the new fashion, the program called SmartBike in Washington D.C. was terminated in January 2011 because of the limited number of stations and low membership. To save the local bike sharing service, a new program called Capital Bike share was launched in D.C. in 2011.

One of the main complaints heard from users of bike sharing systems relates to unavailability of bicycles [5-7], which is the imbalance for the whole systems [1] [8-10]. To prevent the service from failing again, our research will focus on what factors affects bike sharing demand the most and the least. Through this study, our team will analyze the data from 2011 capital Bikeshare program. The data includes the following categories [11]:

season - 1 = spring, 2 = summer, 3 = fall, 4 = winter weather - 1: Clear, Few clouds, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog temp - temperature in Celsiusatemp - sensory temperature in Celsiushumidity - relative humiditywindspeed - wind speedcasual - number of non-registered user rentals initiatedregistered - number of registered user rentals initiatedcount - number of total rentals

To summarize useful information from the above data, perform a topological and statistical analysis using the software R and Python.

II. BACKGROUND

Since there are a total of 12 different categories of variables in the data set, it is hard to analyze. Each category could be a factor that affects the final result, but there are unlikely numbers of data points of wind speed that are zero. It is hard to distinguish whether the data set is zero or a missing collection. So, ignore the effect from the wind speed. Each category except wind speed is being considered in our analysis. Classify the data into two general statistical types: qualitative which are season and weather (Qualitative data is a categorical measurement expressed not in terms of numbers, but rather by means of a natural language description [12]), and quantitative which are temp, atemp, humidity, casual, registered and count (Quantitative data is a numerical measurement expressed not by means of a natural language description, but rather in terms of numbers [12]).

III. METHOD AND RESULT

A. *Data shape and simplify*

First of all, create a 3D graph with x-axis: count, y-axis: temp, z-axis: humidity. (Fig.1).

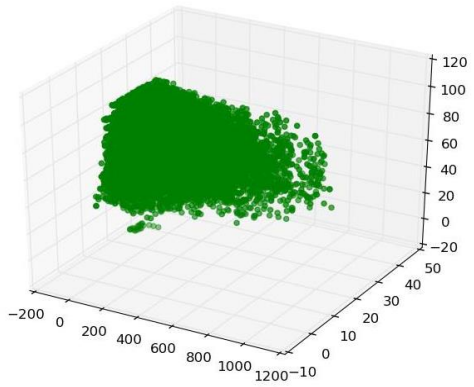
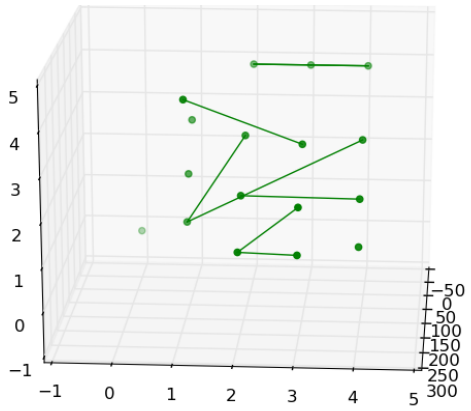


Figure 1. 3-D graph of count, temp and humidity

The shape of the data looks like a Trapezoid, and it's narrow down when the x-axis became bigger. However, there is not much other information which is analyzed from the graph. Then, simplify the data set before any analysis.

In statistic, people normally analyze quantitative data instead of qualitative data. So, the team will use topology knowledge to create a Simplicial Complex (Fig.2) to remove the impact of the qualitative data first. The way to do is, to set the variables (season, weather, sum of the count) as a vector in R^3 space, and then set a fixed length so that if any distance between two vectors is less than the length, set those vectors as complements. Our team narrowed down the data to three complements. Then choose the data point with the highest sum of count which is the point: (Season3, weather1, average count=244). The followings are the graphs with different distance from 10 to 80. With the distance becoming larger, the less complements it represents. Second to last graph shows the three complements that are used.



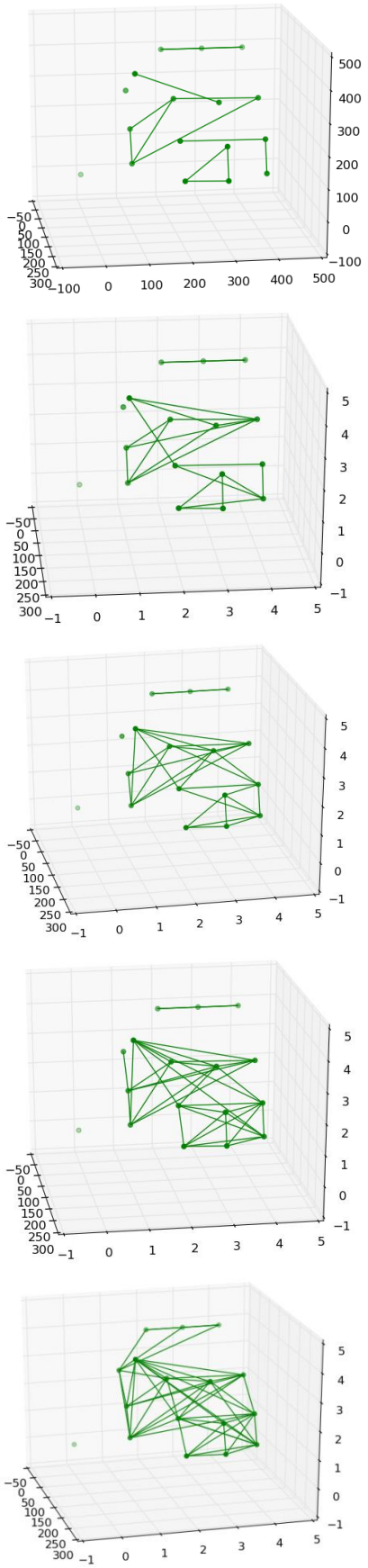
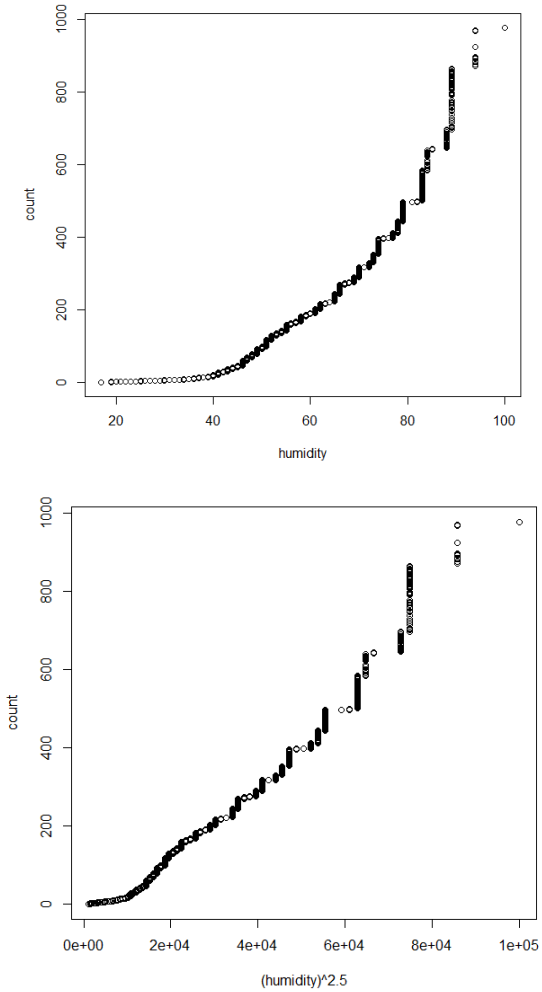


Figure 2. Simplicial complex graph from 10-80

B. *Data transportations*

To analyze the chosen point (quantitative data), since data categories and sample size are quite large, it is hard to fit into one simple mode. Assuming that there is a linear relation between the independent ("variable that is being manipulated in an experiment in order to observe the effect on a dependent variable", x-axis) and dependent ("a variable that is dependent on an independent variable", y-axis) variables, then plug the variable into Multiple Linear Regression model($Yi = \beta0 + \beta1x1i + \beta2x2i + Ei$); since count is equal to causal plus registered, only considering count in this model for right now, so we set count as the dependent variable, and temp, atemp, and humidity as independent variables. However each independent variable may not have accurate linear relation with the dependent variable. So, to create a linear relationship, each independent variable should be normal distribution between with the dependent variable. Data transformation allows users to recreate the independent data set to fit the assumption better, in most of the data transformation, people normally exponentiation or root the data set to change the shape of the graph. To measure the normality between variable, we use QQplot as the method, the more linearity the QQplot is the more normality between variables. The following graphs show the QQplot for temp, atemp, humidity with and without the transformation. According to the Fig.3, the transformation makes the graph much more linear compare to the original one.



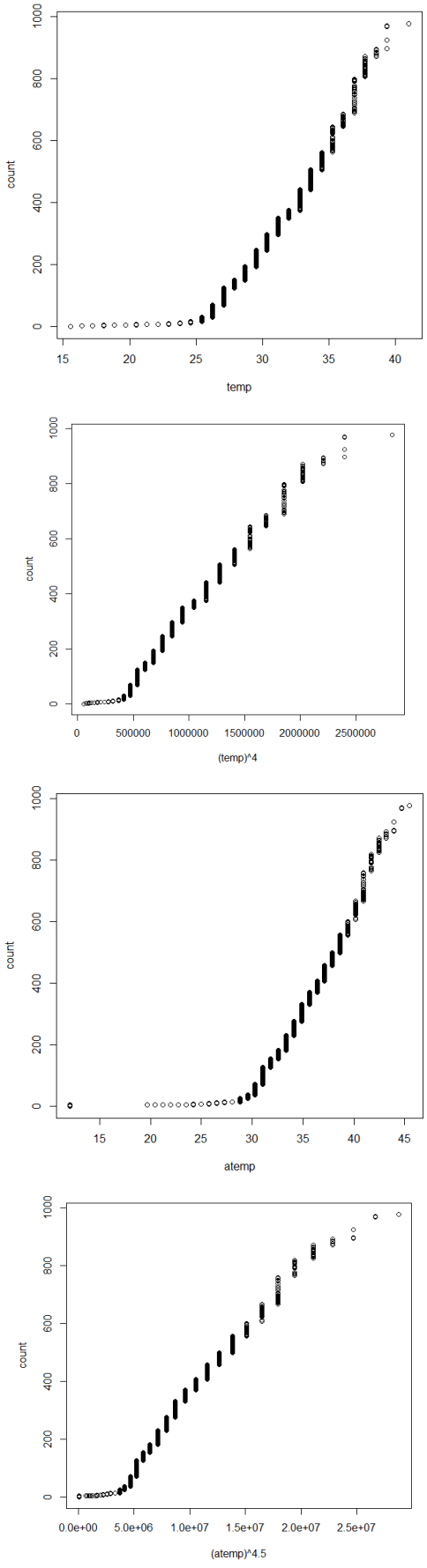


Figure 3. QQ-plots comparison

R code(QQplot):

qqplot(humidity,count) #QQplot between dependent variable and independent variable

qqplot(temp,count) #QQplot between dependent variable and independent variable

qqplot(atemp,count) #QQplot between dependent variable and independent variable

qqplot((humidity)^2.5,count) #QQplot between transformation variable and independent variable

qqplot((temp)^4,count) #QQplot between transformation variable and independent variable

qqplot((atemp)^4.5,count) #QQplot between transformation variable and independent variable

C. Multiple Linear regression model

R code(Multiple Linear regression model):

data=(read.csv(file.choose(), header=TRUE)) # read the data in R

attach(data) # tell R we use the info in data

lm=lm(count~temp^4+atemp^4.5+humidity^2) #create the multiple linear regression between #dependent variable and independent variable

summary(lm)# create the output

R output:

lm(formula = count ~ temp^4 + atemp^4.5 + humidity^2)

Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|) (P-value)	
(Intercept)	402.8126	50.3662	7.998	2.17e-15	***
temp	1.5361	2.2983	0.668	0.5040	
atemp	3.3142	1.7889	1.853	0.0641	.
humidity	-5.1765	0.3087	-16.767	< 2e-16	***

Multiple R-squared: 0.2381, Adjusted R-squared: 0.2369

F-statistic: 200.6 on 3 and 1926 DF, p-value: < 2.2e-16

After done with the data transformation, because we have multiple independent variables, we plug the transformation data into R and creates Multiple Linear Regression model to test the effect. The multiple linear regression is create as:

count= 402.8126+1.5361 temp^4 + 3.3142 atemp^4.5 -5.1765humidity^2

The output not only creates the Multiple Linear Regression model but also produce a hypothesis test for the variable.

H0 : $\beta 1 = \beta 2 = \ldots = \beta k = 0$ The Null hypothesis means the slope for each independent categories are the same and equal to 0.

Ha: not H0 The alternative hypothesis means the case opposite to null hypothesis.

The overall P-value ("the probability that data at least as surprising as the observed sample results would be generated under a model of random chance [13]") from the output is small which means people can reject the null hypothesis. And also from the Estimate it represents the different slope for each category. However, the individual P-value for temp is not small. So, it might mean some of the data do not effect as much as other in this multiple linear regression model.

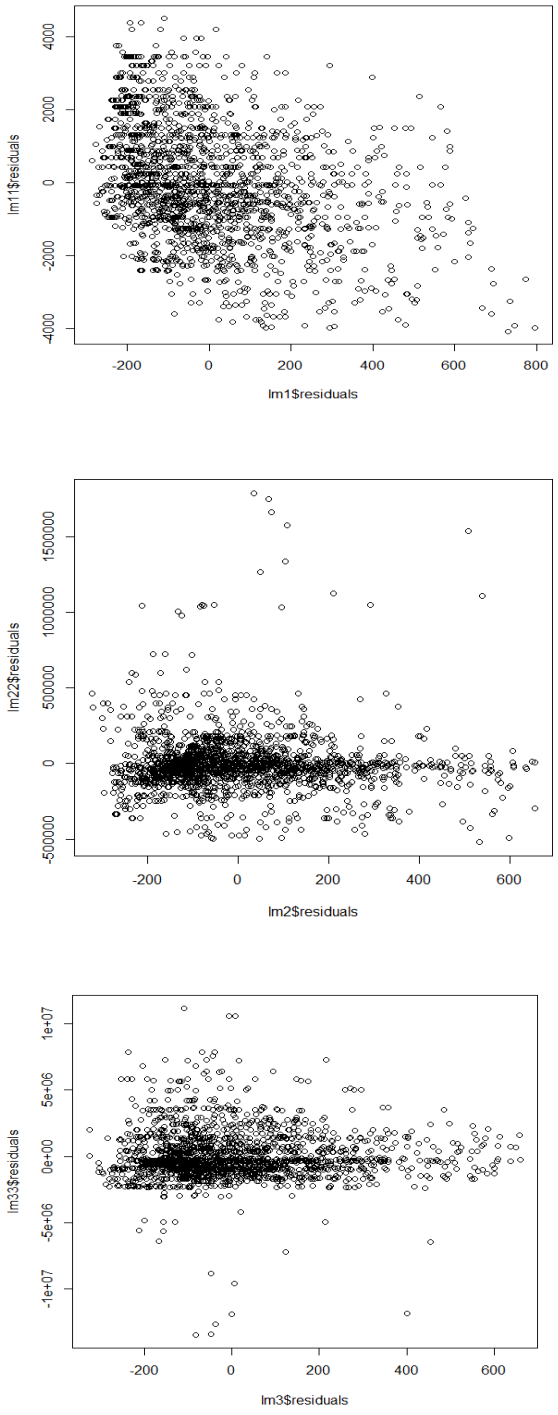


Figure 4. Partial regression plot

D. Partial regression plot

R code:

lm1=lm(count~temp^4+atemp^4.5) # find the multiple linear relation between #count~temp^4+atemp^4.5

lm11=lm(humidity^2~temp^4+atemp^4.5) #find the multiple linear relation between

humidity^2~temp^4+atemp^4.5

lm111=lm(lm1$residuals~lm11$residuals)# find the relation between two groups residuals("an observed value

is the difference between the observed value and the estimated function value.")

summary(lm111)

plot(lm1$residuals,lm11$residuals) #making the partial regression plot

lm2=lm(count~atemp^4.5+humidity^2)

lm22=lm(temp^4~atemp^4.5+humidity^2)

lm222=lm(lm2$residuals~lm22$residuals)

summary(lm222)

plot(lm2$residuals,lm22$residuals)

lm3=lm(count~temp^4+humidity^2)

lm33=lm(atemp^4.5~temp^4+humidity^2)

lm333=lm(lm3$residuals~lm33$residuals)

summary(lm333)

plot(lm3$residuals,lm33$residuals)

To test which variables affect the model the most, our group used another statistics model called partial regression plot ("attempts to show the effect of adding another variable to a model already having one or more independent variables [14]")(the more linear the graph is, the better the factor fitted in the multiple linear regression model is). From Fig.4, it can be observed that the least linearity is the humidity^2. So even though the P-value in the multiple linear regression model is significant small, it is not the variable effecting the relation. And temp and atemp both seem linear according to the partial regression plot. But the p-value for the atemp in the multiple linear regression model is much smaller comparing to temp. So, it can be concluded that the atemp will be the most effective variable under fall season and weather (Clear, Few clouds, partly cloudy).

IV. Discussion

During the research, our team also run other data point in the graph, such as season2 weather1 . Surprisingly, there are different results under different weather and season condition. For example, during the summer, temp and atemp are no longer the most significant factor in the multiple linear regression. The humidity will be the one which affect the most. So, the bike sharing system is actually effect by more on the qualitative data than quantitative data.

How does the weather and season affect the biking sharing system? Is it positive or negative? To solve the problem, we run all the weather, season and count data into Poisson model. The Poisson regression model generate like $\ln(\lambda i) = \beta 0 + \beta 1 x 1 i + \cdots + \beta k x k i$. So it could measures the effect of qualitative data. From the R output below, it sets the season1 weather1 as baseline group, which means other group are compare to the baseline group, and the effect from baseline will be 0. So compare to the baseline group, season2 will increase 85% of the count, season3 will increase 99% of the count, season4 will increase 72% of the count, weather 2 will decrease 12% of the count, weather 3 will decrease 42% of the count, weather 4 will increase 32% of the count.

R code(Poisson):

glm.out=glm(count~season+weather,family=poisson) #create the poisson between count and weather, season.

summary(glm.out)

R output:

Call:

glm(formula = count ~ season + weather, family = poisson)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.822441	0.001847	2610.917	< 2e-16	***
season2	0.616017	0.002214	278.241	< 2e-16	***
season3	0.692518	0.002184	317.142	< 2e-16	***
season4	0.542127	0.002245	241.431	< 2e-16	***
weather2	-0.124203	0.001631	-76.138	< 2e-16	***
weather3	-0.540482	0.003237	-166.977	< 2e-16	***
weather4	0.277425	0.078109	3.552	0.000383	***

Null deviance: 1800567 on 10885 degrees of freedom

Residual deviance: 1640147 on 10879 degrees of freedom

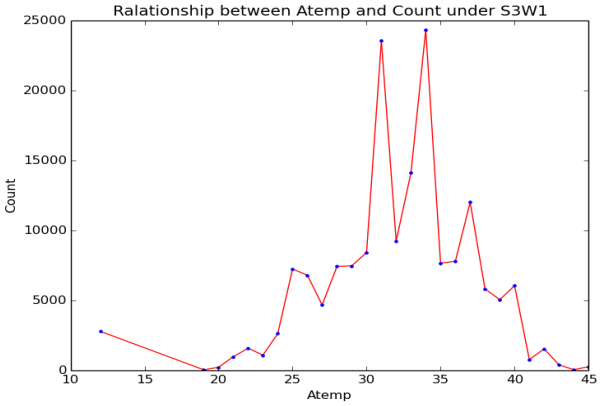


Figure 5. Atemp vs. count

Furthermore, from the previous process, the atemp variable is likely to be the most effective variable for the total count. But how does it exactly affect the bike sharing count. The Fig.5 on the right shows the relation between count and atemp.

V. Check

From the Poisson regression model, the season 3 (Fall) and weather 4 has the most positive effect on the bike sharing system. But since weather 4 is snowy, there are no data point as season3 weather4, the season 3 weather 1 gives the bike sharing system the most positive impact.

VI. Conclusion

Since the project only focuses on what factors impact the bike sharing system, it is concluded what is the most positive effect variables. For conclusion, under the Fall season and weather (Clear, Few clouds, Partly cloudy), the atemp is the most important variable for the bike sharing system. After our further analysis, we found that the atemp, between 30-35℃, is the best sensory temperature for people to rent bicycles.

Acknowledgment

Thank professor Isabel Darcy from Department of Mathematic, University of Iowa for helping us with topology application and programming.

REFERENCES

[1] Luca Di Gaspero, Andrea Rendl, and Tommaso Urli. “Balancing bike sharing systems with constraint programming”. Constraints February 2015.

[2] Maurice H. ter Beek, Alessandro Fantechi, and Stefania Gnesi. “Challenges in Modelling and Analyzing Quantitative Aspects of Bike-Sharing Systems. Leveraging Applications of Formal Methods, Verification and Validation”. Technologies for Mastering Change Lecture Notes in Computer Science Volume 8802, 2014, pp 351-367.

[3] Susan A. Shaheen et al.. "Public Bikesharing in North America During a Period of Rapid Expansion: Understanding Business Models, Industry Trends and User Impacts"(PDF). Mineta Transportation Institure (MTI). Retrieved 2014-11-05. pp. 5.

[4] DeMaio, P.: Bike-sharing: history, impacts,models of provision, and future. J. Public Transp. 12(4), 41–56.

[5] Chung-Cheng Lu. “Robust Multi-period Fleet Allocation Models for Bike-Sharing Systems”. Networks and Spatial Economics August 2013.

[6] Tal Raviv, Michal Tzur, and Iris A. Forma. “Static repositioning in a bike-sharing system: models and solution approaches”. EURO Journal on Transportation and Logistics August 2013, Volume 2, Issue 3, pp 187-229.

[7] Chemla D, Meunier F, and Wolfler Calvo R.. “Bike hiring system: solving the rebalancing problem in the static case”. Discrete Optimization. 2011.

[8] Benchimol M, Benchimol P, Chappert B, Taille ADL, Laroche F, Meunier F, and Robinet L.. “Balancing the stations of a self service ‘‘bike hire’’ system”. RAIRO Operations Research, 45, 37-61.2011.

[9] Marian Rainer-Harbach, Petrina Papazek, Günther R. Raidl, Bin Hu, and Christian Kloimüllner. “PILOT, GRASP, and VNS approaches for the static balancing of bicycle sharing systems”. Journal of Global Optimization April 2014.

[10] Contardo C, Morency C, and Rousseau L-M “Balancing a dynamic public bike-sharing system”. 2012.

[11] Data: Kaggle http://www.kaggle.com/c/bike-sharing-demand/data

[12] Quantitative and Qualitative Data: https://en.wikibooks.org/wiki/Statistics/Different_Types_of_Data/Quantitative_and_Qualitative_Data

[13] P-value: http://en.wikipedia.org/wiki/P-value

[14] Partial regression plot: http://en.wikipedia.org/wiki/Partial_regression_plot.