Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 14
TPCEE 2026

Analysis of Factors Affecting the Rental Volume Based on the Shared Bicycle Data in Washington, D.C

Yifan Zhang *

School of Mathematical Sciences, Beijing University of Posts and Telecommunications, Beijing, China

* Corresponding Author Email: zhangyf2024@bupt.edu.cn

Abstract. The usage of shared bicycles is affected by multiple factors, such as meteorological conditions and time periods. Analyzing the direction and intensity of various variables has practical significance for optimizing operation scheduling and improving system service efficiency. Based on 731 days' daily rental data and matching meteorological records from 2011 to 2012 in the capital bikeshare system in Washington, D.C., this study comprehensively uses descriptive statistics, multiple linear regression, and the random forest model to investigate the impact of temperature, humidity, wind speed, weather conditions, seasons, holidays, and working days on the total number of rentals per day. The multiple linear regression model adjusted R2 to 0.72, the random forest model test set R2 to 0.88, and the root mean square error was 694 times. The results show that temperature is the main positive driving factor, while humidity, rainfall, and wind speed have a significant inhibitory effect on riding intention. The seasonal effect in summer and autumn was significantly higher than that in spring, but there was no significant difference between winter and spring. The year effect significantly reflects the expansion of the system scale. The random forest model further reveals the nonlinear saturation between temperature and rental volume.

Keywords: Bike sharing; rental volume; multiple linear regression; random forest.

1. Introduction

With the characteristics of flexibility, convenience, low-carbon, and environmental protection, bike sharing has become an important part of the urban multi-mode transportation network. The starting and ending points and duration data automatically recorded by the system provide valuable information for quantitative analysis of urban travel laws. Analyzing the impact of different factors on the rental volume not only helps operators to formulate refined scheduling strategies but also provides decision support for traffic management departments to optimize the slow traffic system.

The existing literature has discussed the influencing factors of shared bicycle use behavior from multiple dimensions. At the meteorological level, Liu et al. Found that temperature, humidity, and rainfall have activation effect and threshold effect on turnover by using random forest and interpretable machine learning technology [1]. Gebhart and Nolan's research on Washington, D.C., showed that the number of trips on rainy days decreased by 15% to 30% [2]. Hyland's analysis based on multi-city data further confirmed the universality of weather effects [3]. The empirical analysis in Chicago points out that extreme rainfall can reduce usage by 64% [4]. In terms of built environment and socio-economic dimension, Geipel and others' systematic review identified 39 key spatial factors such as population density, employment density, and proximity of public transport [5]. Wang et al. Further discussed the combined impact of the built environment on usage and riding time [6]. Li et al. Emphasized the importance of price, traffic congestion, and transport capacity supply based on Beijing data [7].

Although the relevant research results are fruitful, the analysis of the effect of multiple factors on the long-term daily data of the Washington D.C. system is still limited. Based on the bike sharing dataset, this study first quantifies the average marginal effect of each factor through multiple linear regression, and then introduces the random forest model to capture the nonlinear correlation between variables. Finally, combined with the policy implications and research limitations of the existing research and

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

discussion results, this paper aims to provide empirical reference for the operation and management of shared bicycles and urban traffic planning.

2. Methods

2.1. Data Source and Description

The research data is from the bike sharing dataset in the UCI machine learning library, and the original data is from the Capital Bike Share system in Washington, D.C., and the corresponding meteorological observations. The data set covers a total of 731 daily records from January 1, 2011, to December 31, 2012. Each sample includes date, season, year, month, holiday identification, working day identification, weather conditions, standardized temperature, standardized body temperature, standardized humidity, standardized wind speed, temporary users, registered users, and total rental times. There is no missing data after preprocessing, so it is suitable to carry out regression analysis and machine learning modeling. This data set is a classic benchmark data set in the field of shared vehicle research and has been widely used in statistical modeling and machine learning algorithm performance verification. The focus of this study is method comparison and factor analysis. The high quality and integrity of classical data sets can provide a reliable basis for this.

2.2. Index Selection and Description

The explained variable is set as the total number of rentals per day. The explanatory variables include the time dimension and the meteorological dimension. The time dimension includes seasons, years, holidays, and working days. Meteorological dimensions include weather conditions (1 sunny, 2 cloudy, 3 light rain or snow, 4 heavy rain or snow, category 4 is merged into category 3 during analysis), standardized temperature, standardized humidity, and standardized wind speed. The classification variables are processed by one-hot encoding.

2.3. Method Introduction

The data analysis process consists of three parts. Firstly, the time trend and temperature correlation of the rental volume are visually explored through the histogram and scatter chart, which visually shows the seasonal fluctuation law of the data and the potential nonlinear relationship between variables, and provides a preliminary basis for the subsequent model setting. Secondly, the Pearson correlation coefficient is calculated to test the linear correlation between the explanatory variables and the rental volume, and the multicollinearity between variables is checked to ensure the stability and reliability of the regression coefficient estimation. Then, a multiple linear regression model was constructed, taking the total number of rentals as the dependent variable, including season, year, holidays, working days, weather conditions, temperature, humidity, and wind speed as independent variables, in which season and weather conditions were transformed into virtual variables, and spring and sunny weather were taken as the reference group. The advantage of multiple linear regression is that it can control the interference of other variables, give the independent marginal contribution of each factor to the rental volume, and the meaning of the coefficient is intuitive and easy to explain, which is suitable for answering such questions as "the influence direction and average intensity of each factor". The model expression is:

$$\begin{aligned} \mathrm{cnt} = \beta_0 + \beta_1 \cdot \mathrm{season_summer} + \beta_2 \cdot \mathrm{season_fall} + \beta_3 \cdot \mathrm{season_winter} + \beta_4 \cdot \mathrm{yr} + \\ \beta_5 \cdot \mathrm{holiday} + \beta_6 \cdot \mathrm{workingday} + \beta_7 \cdot \mathrm{weathersit_cloudy} + \beta_8 \cdot \mathrm{weathersit_rain} + \\ \beta_9 \cdot \mathrm{temp} + \beta_{10} \cdot \mathrm{hum} + \beta_{11} \cdot \mathrm{windspeed} + \varepsilon \end{aligned} \quad (1)$$

Where β_0 is the intercept term, β_1 to β_{11} are the regression coefficients of the respective variables, and ε is the random error term. The advantage of multiple linear regression is that it can control the interference of other variables, give the independent marginal contribution of each factor to the rental volume, and the meaning of the coefficient is intuitive and easy to explain.

In order to make up for the deficiency of the linear model in capturing nonlinear relationships, the random forest algorithm is further used for modeling. By integrating a large number of decision trees and random sampling of characteristics, a random forest can automatically depict the complex nonlinear dependence between independent variables and dependent variables, and output the importance ranking of variables, providing another perspective for identifying key drivers. In the analysis process, the data set is divided into a training set and a test set according to the ratio of 7:3. The number and maximum depth of grid search optimization trees and other super parameters are used to evaluate the performance of the model with root mean square error and determination coefficient.

3. Results and Discussion

3.1. Visual Analysis

Fig. 1 shows the evolution trend of the average monthly rental volume from 2011 to 2012. It can be seen that the usage shows a significant seasonal cycle, with the highest rental in summer and the lowest in winter, which is consistent with the characteristics of the temperature saturation effect proposed by Liu et al. [1].

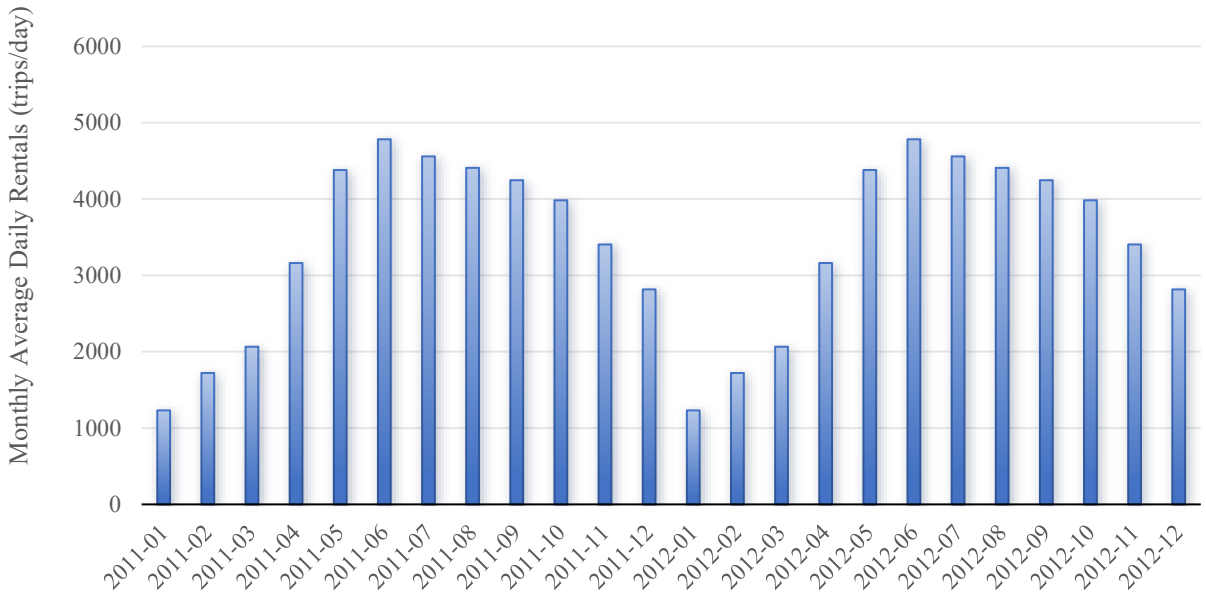

Fig. 1 Monthly average daily rentals in Washington, D.C. (2011–2012) (Photo credit: Original)

Fig. 2 shows the relationship between standardization temperature and total rental volume. There is a positive correlation between the two as a whole, but there is a slight drop in areas where the standardized temperature exceeds 0.8, suggesting that extremely high temperatures may inhibit travel willingness, echoing the research findings that vulnerable communities are less sensitive to extremely high temperatures [4].

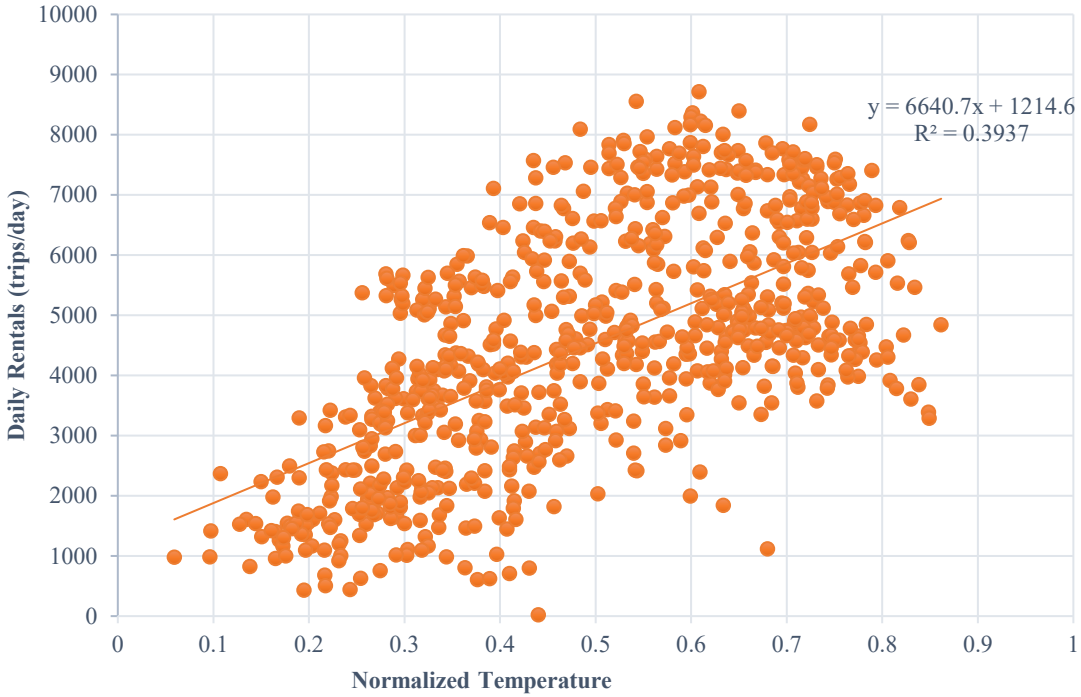

Fig. 2 Scatter plot of normalized temperature vs. total daily rentals (Photo credit: Original)

3.2. Correlation Analysis

Pearson correlation analysis showed that the total rental volume was positively correlated with temperature ($r=0.63$, $p<0.01$) and year ($r=0.57$, $p<0.01$), and negatively correlated with humidity ($r=-0.35$, $p<0.01$) and wind speed ($r=-0.23$, $p<0.01$). There was a positive correlation between summer and rental volume ($r=0.42$), and a negative correlation between winter and rental volume ($r=-0.51$). There was a weak negative correlation between holidays and rental volume ($r=-0.12$), while there was no significant linear correlation between weekdays. The correlation coefficients among the explanatory variables were lower than 0.7, indicating that there was no serious multicollinearity. These correlations are basically consistent with the existing research conclusions and provide a preliminary basis for the subsequent model construction.

3.3. Regression Model Results

The multiple linear regression model was significant as a whole ($f=189.4$, $p<0.001$), and the adjusted R^2 was 0.72. The regression coefficient and significance test results of each variable are listed in Table 1.

Table 1. Multiple linear regression results

Variable	Coefficient	Standard Error	T Value	P Value
Intercept	1280.5	215.3	5.95	< 0.001***
Summer	1120.3	140.2	7.99	< 0.001***
Autumn	850.6	135.8	6.26	< 0.001***
Winter	-210.4	142.1	-1.48	0.139-
Year	2034.7	92.5	22.00	< 0.001***
Holiday	-510.2	180.6	-2.83	0.005**
Working Day	120.3	102.4	1.17	0.241-
Cloudy Weather	-350.1	110.5	-3.17	0.002**
Rain/Snow Weather	-1520.4	180.2	-8.44	< 0.001***
Temperature	4100.2	380.1	10.79	< 0.001***
Humidity	-2150.5	320.4	-6.71	< 0.001***
Windspeed	-980.3	410.2	-2.39	0.017*

Note: * * *: P<0.001 * *: P<0.01; *:<0.05; p ≥ 0.05; -It means not significant. The season is spring, as the reference group, and the weather is sunny, as the reference group.

The regression results show that temperature has the strongest positive effect on the rental volume, and humidity and wind speed have significant negative effects. Compared with spring, the number of rentals in summer and autumn was 1120 and 851 times higher, respectively, and the difference between winter and spring was not statistically significant. The year effect is extremely significant, and the average daily rental volume in 2012 is about 2035 times higher than that in 2011. The rental volume during holidays is 510 times less than that during non-holidays on average. The weekday effect is not significant, suggesting that bike sharing has surpassed the single function of commuting tools in Washington, D.C. In terms of weather conditions, the average number of cloudy days decreased by 350 times compared with sunny days, while the number of rainy and snowy days decreased by 1520 times. It is worth noting that the coefficient of winter variables is not significantly negative after taking into account temperature and other meteorological factors, which reflects the value of multiple regression to control mixed variables: Winter itself has no statistically significant independent inhibitory effect on riding demand.

3.4. Results of the Random Forest Model

The determination coefficient of the random forest model on the test set is 0.88, and the root mean square error is 694. The prediction performance of the random forest model is significantly improved compared with that of the linear model, indicating that there is a nonlinear model in the data that is difficult to fully fit into the linear structure. The ranking of characteristic importance showed that temperature, year, and humidity ranked the top three, which was consistent with the core findings of regression analysis. The positive effect of temperature on rental volume slowed down and stabilized after the standardized temperature exceeded 0.7, confirming the temperature threshold effect reported in the literature [1]. The relationship between humidity and rental volume also showed an approximate inverse S-shaped relationship. When the standardized humidity exceeded 0.6, the rental volume accelerated to decline, further supporting the hypothesis of a humidity threshold.

4. Discussion

Through the mutual confirmation of multiple linear regression and the random forest model, this study identified temperature, year, humidity, and rainfall as the key factors affecting the rental volume. The linear model results show that the selected variables have good explanatory power, but about 28% of the variation can not be captured by the current factors. The random forest model improves the fitting degree to 0.88 by fitting the nonlinear relationship, highlighting the importance of the threshold effect between meteorological variables and rental volume.

However, there are still some limitations. First of all, the data only covers two years, and the space is limited to Washington, D.C. conclusion the promotion to other cities or different climate zones needs careful verification, and there may be significant differences in the use patterns of shared bicycles between different cities [8]. Secondly, due to the information provided by the data set, it failed to include important variables that may affect riding choices, such as the built environment, public transport accessibility, road infrastructure, fuel prices, and shared bicycle pricing strategies [9, 10]. Third, the model does not consider the interactive effects of weekends and holidays, nor does it model different types of users (registered users and temporary users), which may mask the heterogeneous travel demand. Fourth, although the random forest improves the prediction ability, its black box attribute weakens the intuition of mechanism interpretation to a certain extent.

In view of the above shortcomings, future research can be expanded from the following aspects: introducing spatial variables such as land use mixing degree, subway station distribution, bicycle lane density, and building a multi-level model to separate individual and environmental effects; The time series method is used to describe the autocorrelation structure of rental volume and the lag weather effect; Distinguish between registered users and temporary users, and investigate the response

differences of different groups to meteorological and time factors; Using interpretable machine learning technology such as shap value, the model transparency and policy availability are enhanced while maintaining the prediction accuracy.

5. Conclusion

Based on the daily rental and meteorological data of the shared bicycle system in Washington, D.C., for two consecutive years, this study systematically investigated the effects of weather, season, and time factors on the rental volume by using multiple linear regression and the random forest model. The main conclusions are as follows: Firstly, temperature is the primary positive factor driving the increase in rental volume, while humidity, rainfall, and wind speed significantly inhibit the willingness to ride. Second, the rental volume in summer and autumn is significantly higher than that in spring; the winter effect is not significant, and the year effect reflects the rapid expansion of the system scale. Thirdly, the random forest model confirms that the impact of temperature and humidity on the rental volume has nonlinear threshold characteristics, which provides a quantitative basis for operators to set dynamic scheduling strategies based on meteorological indicators. Fourth, the number of rentals during holidays decreased significantly, but the weekday effect was not significant, indicating that the use of shared bicycles in Washington, D.C. is no longer limited to commuting. The research results can provide empirical support for bike sharing enterprises to optimize vehicle distribution, formulate weather response plans, and improve the slow traffic system for urban traffic management departments. Fifthly, the discussion part emphasizes the limitations of the research in terms of data space-time range, variable completeness and model interpretability, and points out the direction for the introduction of spatial analysis and interpretable machine learning in the future.

References

[1] Liu P, Pan Z, Fan Z, et al. The impact of weather on shared bikes. Applied Sciences, 2025, 15(17): 9834.

[2] Gebhart K, Nolan R. Weather and bike-share trips: a case study of the Capital Bikeshare system in Washington, D.C. Transport Findings, 2022.

[3] Hyland M, Hong Z, Pinto H. The impact of weather on bikeshare trips: a multi-city analysis. International Journal of Sustainable Transportation, 2018, 12(7): 511-521.

[4] Ermagun A, Zhang E. Mobility of disadvantaged communities exhibits lower sensitivity to extreme weather. Travel Behaviour and Society, 2025, 38: 100812.

[5] Geipel M, Martínez Rico B, Büttner B, et al. Spatial factors associated with usage of different on-demand elements within mobility hubs: a systematic literature review. Transport Reviews, 2024, 44(3): 345-367.

[6] Wang K, Chen Y J. Joint analysis of the impacts of built environment on bike share usage and trip duration. Journal of Transport Geography, 2020, 82: 102631.

[7] Li L, Liu Y, Song Y. Factors affecting bike-sharing behaviour in Beijing: price, traffic congestion, and supply chain. Annals of Operations Research, 2025, 349(2): 626-645.

[8] Noland R B, Smart M J, Guo Z. Bikeshare trip generation in New York City. Transportation Research Part A, 2016, 94: 164-181.

[9] El-Assi W, Mahmoud M S, Habib K N. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. Transportation, 2017, 44(3): 589-613.

[10] Campbell K B, Brakewood C. Sharing riders: how bikeshare ridership is influenced by transit proximity. Journal of Transport and Land Use, 2017, 10(1): 251-270.