Design and implementation of analysis and visualization of shared bicycle information

Haiyan Ji, Hui Xu, Cheng Zhou, Jingyi Zhao
School of Information Science & Technology, Nantong University, Nontong 226019, China
Corresponding author and e-mail: Hui xu, xu.h@ntu.edu.cn

Abstract. While the shared bicycle without piles solves the problem of people’s "last mile" travel, its free mobility also brings the embarrassing situation of "one car is idle, the other is looking for it". In this paper, data analysis methods are discussed to reveal the different flow patterns of shared bicycles usage and mobility variation with seasons, weather, time periods, working days, etc; predict method is given for the number of shared bicycle rides based on random forest algorithm. Mastering the distribution and riding rules of shared bicycles can help improve the efficiency of the use of bicycles, better meet the riding needs of users, and provide a data-based basis for regulating the regional regulation of shared bicycles.

1. Introduction
Under the influence of big data, cloud computing and mobile internet technology, the innovative practice of "mobile internet + travel" was carried out in 2016 [1]. Bike sharing is the product of this innovation, and its appearance helps to alleviate the difficulty of traveling the "last kilometers"[2]. With mobile payment and GPS technology as its strong support, shared bicycles have developed rapidly as a new type of green travel vehicle. It has long become an urban landscape and gradually changes the urban spatial layout. The pile-less design also realizes the advantages of shared bicycles with any use and parking. However, pileless parking also leads to scattered and disordered shared bicycles. If you cannot know the riding characteristics and layout rules of shared bicycles, you cannot effectively regulate them, which makes bicycle sharing bring great convenience to people's lives, but also brings unprecedented pressure and challenges to urban traffic management and facility planning.

The flow of shared bicycles seems to be irregular, but in fact there are rules to follow [3]. This paper studies the shared bicycle riding rules hidden after riding data. These laws can provide travel options for rail transit connections, and provide references for standardizing shared bicycle development programs and encouraging green travel methods.

2. Related work
2.1. Development environment
The programming language is Python. Python's rich third-party libraries provide great convenience [4] for solving problems such as data analysis and artificial intelligence. The third-party libraries involved in this paper are: numpy, pandas, maiplotlib, seaborn and sklearnn.

- Numpy, provide array support.
• Matplotlib, used to draw graphics and visualize data information.
• Pandas, contains data structure and operation tools to help us clean and analyze data.
• Sklearn, a machine learning library, uses it to predict cycling information for shared bicycles.

2.2. Random forest algorithm
Random forest algorithm is [6] a supervised machine learning algorithm, a classifier that uses multiple trees to train and predict samples, and the category of its output is determined by the mode of the category that the individual tree outputs. Random forest construction has two aspects: random selection of data and random selection of features to be selected.

2.3. Data sources
Data set is provided by the Kaggle with information on time, climate and amount of riding. Data set training set has a total of 8600 data, each with an hour interval, covering user cycling data on 1~15 each month in 2011 and 2012. The sample data is shown in Figure 1.

![Figure 1. Sample of data set.](image)

The meaning of each field is shown in Table 1.

| datetime | season | holiday | workingday | weather | temp | atemp | humidity | windspeed | casual | registered | count |
|----------|--------|---------|------------|---------|------|-------|----------|----------|--------|------------|-------|
| 2011/1/10 00:00 | 1 | 0 | 1 | 1 | 4.92 | 6.06 | 50 | 19.0012 | 2 | 3 |
| 2011/1/10 1:00 | 1 | 0 | 1 | 1 | 4.92 | 6.06 | 50 | 19.0012 | 1 | 0 |
| 2011/1/10 2:00 | 1 | 0 | 1 | 1 | 4.92 | 6.06 | 50 | 15.0013 | 0 | 3 |
| 2011/1/10 3:00 | 1 | 0 | 1 | 1 | 4.92 | 6.06 | 50 | 15.0013 | 0 | 1 |
| 2011/1/10 4:00 | 1 | 0 | 1 | 1 | 4.1 | 6.06 | 54 | 8.9981 | 1 | 2 |
| 2011/1/10 5:00 | 1 | 0 | 1 | 1 | 4.1 | 5.305 | 54 | 16.9979 | 0 | 3 |
| 2011/1/10 6:00 | 1 | 0 | 1 | 1 | 4.92 | 6.06 | 50 | 19.0012 | 0 | 31 |
| 2011/1/10 7:00 | 1 | 0 | 1 | 1 | 4.92 | 6.06 | 50 | 15.0013 | 2 | 75 |
| 2011/1/10 8:00 | 1 | 0 | 1 | 2 | 4.92 | 6.06 | 50 | 19.0012 | 4 | 184 |
| 2011/1/10 9:00 | 1 | 0 | 1 | 2 | 5.74 | 6.06 | 50 | 16.9979 | 2 | 92 |
| 2011/1/10 10:00 | 1 | 0 | 1 | 2 | 5.74 | 6.06 | 50 | 19.9995 | 0 | 31 |
| 2011/1/10 11:00 | 1 | 0 | 1 | 2 | 6.56 | 6.82 | 47 | 19.0012 | 2 | 28 |
| 2011/1/10 12:00 | 1 | 0 | 1 | 2 | 8.2 | 9.09 | 40 | 19.0012 | 5 | 47 |
| 2011/1/10 13:00 | 1 | 0 | 1 | 2 | 8.2 | 9.09 | 40 | 19.0012 | 4 | 50 |
| 2011/1/10 14:00 | 1 | 0 | 1 | 2 | 8.2 | 9.85 | 40 | 15.0013 | 0 | 47 |
| 2011/1/10 15:00 | 1 | 0 | 1 | 2 | 8.2 | 9.85 | 40 | 15.0013 | 2 | 43 |
| 2011/1/10 16:00 | 1 | 0 | 1 | 1 | 8.2 | 10.605 | 40 | 8.9981 | 4 | 70 |
| 2011/1/10 17:00 | 1 | 0 | 1 | 1 | 8.2 | 11.365 | 40 | 7.0015 | 4 | 174 |
| 2011/1/10 18:00 | 1 | 0 | 1 | 1 | 8.2 | 9.65 | 40 | 16.0013 | 1 | 154 |
| 2011/1/10 19:00 | 1 | 0 | 1 | 1 | 8.2 | 8.305 | 47 | 11.0014 | 3 | 92 |
| 2011/1/10 20:00 | 1 | 0 | 1 | 1 | 8.2 | 8.305 | 50 | 11.0014 | 1 | 73 |
| 2011/1/10 21:00 | 1 | 0 | 1 | 1 | 5.74 | 6.82 | 50 | 12.998 | 1 | 37 |

**Table 1.** The meaning of each field in data set.

- **datetime**: Date and time of data recording.
- **season**: 1.2,3 and 4 correspond to spring, summer, autumn and winter.
- **holiday**: Is it a holiday today.
- **workingday**: Is today a working day.
- **weather**: 1 sunny, cloudy; 2 foggy; 3 light rain, light snow; 4 heavy rain, hail, blizzard.
- **temp**: Centigrade temperature.
- **atemp**: Body temperature in centigrade.
- **humidity**: Humidity.
- **windspeed**: Wind speed.
- **casual**: Vehicles for non-registered users.
- **registered**: Vehicles for Registered user.
- **count**: Total vehicle use.
3. Analysis and visualization of shared cycling information

3.1. Analysis of temporal and climatic characteristic

The effects of working day or not, season, weather, temperature and humidity on the use of shared bikes are shown in Figure 2.

3.1.1. Ride pattern of working days or not. Figure 2 a) shows the number of cyclists per hour in 24 hours a day during working and non-working days. It can be seen from the figure that the peak period of working days is very obviously, from 7:00 am to 8:00 am and 17:00 pm to 19:00 pm, which is exactly the peak period of commuting. During the day, the number of cyclists changes are not obvious, in the middle of the day, The number of cyclists in the evening gradually decreased. The time distribution of non-working days was average, the characteristic curve was relatively mild, the travel began to grow slowly after 8:00, and the users of cycling gradually decreased after 17:00 in the afternoon. It could be seen that the demand for working days and non-working days is completely different, the demand for working days is mainly commuting, the demand for non-working days is mainly living, and the amount of riding reaches the maximum at noon.

3.1.2. Impact of seasons on vehicles. Figure 2 b) The number of cyclists per hour per season is given based on statistics. The figure shows that there are fewer cyclists in spring than in other seasons. The overall number of cyclists during the day is relatively stable, Around 250. Renting bikes concentrated in summer and autumn, the change of vehicles at all times has little to do with seasons, even in winter relative to other times, 7:00 am-8:00 am and 17:00 pm-19:00 pm are still rush hours, and reached the time peak at 8:00 am and 17:00 pm respectively. In all seasons, Because of working hours, the number of cyclists users during the day is low between 9:00 am and 17:00 pm.

3.1.3. Impact of weather on vehicles. Figure 2 c) the number of cyclists in different weather months is given based on the statistics. From the figure, we can see that the temperature is more suitable from May to October, and sunny days and cloudy are prosperous weather of rental bikes, so the number of cyclists is rising. The number of cyclists in foggy weather has not been greatly affected, and the

![Figure 2](Image)
overall number of cyclists in each month of light rain is on the low side. Because of safety reasons, heavy rain and heavy snow weather almost no one rents bikes.

3.1.4. Impact of temperature and humidity on vehicles. Figure 2) The variation of the number of cyclists in different temperature and humidity intervals is given based on statistics. The figure shows: in the temperature between 10 to 40 degrees, humidity between 40 to 80 conditions, there are more people renting bicycles. Temperature below 10 and above 30 degrees, the number of cyclists is relatively small, indicating that too hot and too cold will inhibit the demand for bicycle rental. Further analysis shows that when the summer is dry, the number of users renting bicycles is more, and in winter, when the humidity is moderate, the number of cyclists is also large. The autumn weather is cool and the temperature and humidity are suitable, which is a period of strong demand for bicycles rental. When the humidity is high, the car rental volume drops significantly, indicating that people prefer cycling in dry weather.

3.2. Prediction of riding quantity based on Random Forest algorithm
The columns related to shared bicycle usage are count (total), registered (registered users’ usage) and casual (non-registered users' usage), count = registered + casual.

3.2.1. Feature correlation analysis. For the training set of data set, uses the dataframe.corr () function in the pandas library to analyze the data correlation between any two columns. The correlation calculation method selects the default method parameter pearson (Pearson correlation coefficient). Correlation value is between -1 and 1, a positive number meaning a positive correlation, a negative number meaning a negative correlation, and a larger absolute value means a stronger correlation. The code is as follows:

Correlation = train_datafram [ ["season","temp", "atemp", "humidity", "windspeed", "casual", "holiday", "workingday", "weather", "registered", "count" ]]. corr()

Thermal diagram visualization of the calculated results of correlation coefficients, as shown in Figure 3.

Figure 3. Thermal map of correlation coefficient.

The characteristics of casual and registered determined by Figure 3 are as follows:
Casual selection is characterized by: hour, temp, year, month, season, workingday, windspeed, week, weather, humidity, holiday.
Registered selection is characterized by: hour, temp, year, month, season, workingday, windspeed, week, weather, humidity, holiday.
3.2.2. Description and implementation of algorithm. The Random Forest algorithm is used to predict the amount of riding. The steps are as follows:

Step 1: combining the test set and the training set, the features with high correlation with the target value are preliminarily selected (as described above).

Step 2: using One-Hot coding to extract class-specific features that can not directly participate in operations (e.g. year, month, day, hour, week, season, weather).

Step 3: extracting the features that are ultimately used for prediction and the log-transformed label values (the number of vehicles used by non-registered users and the number of vehicles used by registered users, respectively).

Step 4: splitting training set and test set.

Step 5: training random forests based on features and labels extracted by Step 3, this paper uses 1000 trees as a training subset to find the most suitable labels.

Step 6: inputing test set sample, the prediction result of each decision tree is its leaf node value. In this paper, the average value of random forest leaf node is taken as the prediction value.

predicted results are shown in Figure 4 a). comparing the prediction results with the training set (figure 1), it is found that the prediction results are basically accurate within the error allowable range.

![Figure 4](image)

Figure 4. The prediction result of number of shared bicycles.

4. Conclusions
Sharing bicycle is flexible and convenient, low-carbon and environmentally friendly, economical and energy-saving, and has good accessibility. It has gradually become the first choice to solve the city "last kilometer" travel. Through the visual analysis of shared cycling data, It is revealed that there are different characteristics of cycling time and space under the influence of time, climate and season. These characteristics can provide a reliable and valuable reference for rational planning of shared bicycle delivery and bicycle scheduling, and promote the healthy development of public bicycle system under the mode of sharing economy to a certain extent.

References
[1] YUE Yu-jun, HU Han-hui. The Research on the Governance Countermeasure of Shared Bicycle based on Multi-theoretical Perspective [J]. Journal of Technical Economics & Management, 2019(02):86-91.
[2] Labadi K, Benarbia T, Barbot J P, et al. Stochastic petri net modeling, simulation and analysis of public bicycle sharing systems[J]. IEEE Transactions on Automation Science & Engineering, 2017, 12(4):1380-1395.
[3] Lifang Den, Yonghong Xie, Dingxi Huan. Bicycle-sharing Facility Planning Base On Riding Spatio-temporal Data/Deng Lifan, Xie Yonghong, Huang DingxiPlanner2017,33(10):82-88.
[4] http://pypi.org
[5] An Jin, Cheng Cheng, Shuhua Song, et. al. Regional query of area data based on geohash[J]. Geography and Geo-Information Science, 2013, 29(05): 31-35
[6] Xiaoyi Zhou, Pan Lu, Zijian Zheng, Denver Tolliver, Amin Keramati. Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree [J]. Reliability Engineering and System Safety, 2020, 200.