Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 8847752, 13 pages
https://doi.org/10.1155/2020/8847752

Research Article

Machine Learning Approach to Quantity Management for Long-Term Sustainable Development of Dockless Public Bike: Case of Shenzhen in China

Qingfeng Zhou,[1,2] **Chun Janice Wong**,[1] **and Xian Su**[3]

[1]*Harbin Institute of Technology, Shenzhen, Guangdong 518055, China*
[2]*Shenzhen Key Laboratory of Urban Planning and Decision Making, Shenzhen, Guangdong 518055, China*
[3]*China Resources Land Guangxi Co, Nanning, Guangxi 530000, China*

Correspondence should be addressed to Chun Janice Wong; janicewong@hit.edu.cn

Received 21 July 2020; Revised 25 October 2020; Accepted 12 November 2020; Published 28 November 2020

Academic Editor: Petr Dolezel

Since the number of bicycles is critical to the sustainable development of dockless PBS, this research practiced the introduction of a machine learning approach to quantity management using OFO bike operation data in Shenzhen. First, two clustering algorithms were used to identify the bicycle gathering area, and the available bike number and coefficient of available bike number variation were analyzed in each bicycle gathering area's type. Second, five classification algorithms were compared in the accuracy of distinguishing the type of bicycle gathering areas using 25 impact factors. Finally, the application of the knowledge gained from the existing dockless bicycle operation data to guide the number planning and management of public bicycles was explored. We found the following. (1) There were 492 OFO bicycle gathering areas that can be divided into four types: high inefficient, normal inefficient, high efficient, and normal efficient. The high inefficient and normal inefficient areas gathered about 110,000 bicycles with low usage. (2) More types of bicycle gathering area will affect the accuracy of the classification algorithm. The random forest classification had the best performance in identifying bicycle gathering area types in five classification algorithms with an accuracy of more than 75%. (3) There were obvious differences in the characteristics of 25 impact factors in four types of bicycle gathering areas. It is feasible to use these factors to predict area type to optimize the number of available bicycles, reduce operating costs, and improve utilization efficiency. This work helps operators and government understand the characteristics of dockless PBS and contributes to promoting long-term sustainable development of the system through a machine learning approach.

1. Introduction

The public bike system (PBS) also called a bicycle sharing system (BSS), which was born in 1965 in Europe, has been developed for three generations [1]. PBS is economical, eco-friendly, healthy, more equitable, produces ultralow carbon emissions, and has rapidly emerged in many cities all over the world [2]. Since 2016, a relatively new model of PBS, known as the free-float bike sharing system, has increasingly gained its popularity. The FFBS is based on the mobile app and GPS which eliminates stations and docks (also called dockless bike). Passengers can easily pick up and drop off a bike anywhere using their cell phones. This system is quite spread nowadays through enterprises as OFO and Mobike since early 2016 in China. Dockless PBS brings new experiences and conveniences as well as some problems, and an important issue is to consider the number of bicycles available. It has two sides:(1) assuming that the surrounding roads are suitable for cycling when a large number of dockless bicycles are concentrated in an area with a low cost to use, we can think it is providing "adequate" bike supply, which can help us fully understand the bike demand in this area; (2) in fact, if the number of available bicycles is too large, it will cause a series of waste. Many problems are related to the number of shared bicycles especially for the dockless PBS which is an important issue to be considered. But it is seldom involved in the existing research. The number of available bicycles is the core indicator. Excessive

bicycles can affect the cost and efficiency of operation, which is not conducive to the long-term sustainability of the system. The government and scholars have paid more and more attention to the question of how to rationally develop dockless PBS in the city.

Computational intelligence, such as artificial neural networks, fuzzy systems, and evolutionary computing, has achieved significant results in modeling, learning, and search and optimization problems for smart city applications [3, 4]. The characteristics of machine learning make it attractive for analyzing smart city data with complex nature [5], such as modes (streams, time series, images, videos, and texts), large amounts (continuous data generated by millions of sensing devices), space-time dependence, etc. Researchers in smart cities have applied machine learning in many areas, such as urban human mobility [6], public space utilization [7], and public bus charging station placement [8]. Since the number of bicycles is critical to the sustainable development of dockless PBS, this research practiced the introduction of a machine learning approach to quantity management. Four issues are discussed from the existing shared bicycle operation data. (1) How to identify the gathering area of dockless shared bicycles? (2) How to measure the number of bicycles and activity characteristics in the bicycle gathering area? (3) What are the differences between classification algorithms in predicting the types of bicycle gathering areas? (4) How to use activity pattern to guide dockless PBS rationally develop in the city? In this study, first, two clustering algorithms were used to identify the bicycle gathering area, and the available bike number and coefficient of available bike number variation were analyzed in each bicycle gathering area's type. Second, five classification algorithms were compared in the accuracy of distinguishing the type of bicycle gathering areas using impact factors. Finally, the application of the knowledge gained from the existing dockless bicycle operation data to guide the number planning and management of public bicycles was explored.

The rest of this paper is organized as follows. Section 2 presents a literature review on the systems perspective of public bike research. Section 3 introduces the indicators and methods used. Section 4 briefly describes bicycle operation data and influencing variables. In Section 5, we discuss the bicycle gathering area type's recognition, prediction, and application. Finally, Section 6 summarizes the results of this study and provides direction in future studies.

2. Previous Work

PBS is involved in many areas of research, and it is broadly based on two perspectives: user perspective and systems perspective [9]. In this study, we only focus on a systems perspective according to the goals.

2.1. Bike Sharing Rebalance. For PBS, the lack of resources is the major issue: a user can arrive at a station that has no bike available or wants to return her bike at a station with no empty spot. Based on the practical usage, several studies focused to deal with public bike rebalancing problem using intelligent algorithms. Fricker and Gast [10] proposed a stochastic model of homogeneous PBSs to study the effect of users' random choices on the number of problematic stations. They also computed the rate of which bikes must be redistributed by trucks to ensure a given quality of service. You et al. [11] provided an integrated model to resolve the problems of fleet sizing, empty-resource repositioning, and vehicle routing for bike transfer in multiple station systems. O'Mahony and Shmoys [12] tackled the problem of rebalancing PBS during rush hour. An optimization problem whose goal is to plan truck routes to make PBS as balanced as possible in night shift was studied, and novel methods were developed for optimizing rebalancing resources. Chen et al. [13] addressed the layout planning of public bicycle system within the attracted scope of a metro station. Locations of different PBS service stations and the optimal route options for the implementation of the redistributing strategy were considered. Lozano et al. [14] proposed a multiagent model that provides visualization and prediction tools for PBS.

2.2. Bike Demand Estimation. These studies examine the influence of PBS infrastructure, transportation network infrastructure, land use and urban form, meteorological data, and temporal characteristics on PBS usage. Faghih-Imani et al. [15] collected station-level occupancy data and then transformed station occupancy snapshot data into station-level customer arrivals and departures. They developed a mixed linear model to estimate the influence of bicycle infrastructure, sociodemographic characteristics, and land-use characteristics on customer arrivals and departures. In the work of Krykewycz et al. [16], various demographic, land use, and infrastructure factors understood to be favorable for bike share usage were spatially analyzed to define a primary market area. El-Assi et al. [17] investigated the effects of weather, socioeconomic and demographic factors, and land use and the built environment on bicycle share ridership. A regression analysis was performed on three different levels. Hampshire and Marla [18] employed a panel regression model to explain the factors affecting the bike sharing trip generation and attraction in the presence of unobserved spatial and temporal variables. The data used included PBS's usage data in Barcelona and Seville, nine census demographic data, and the location of points of interest (POIs). Zhang et al. [19] employed a multiple linear regression model to examine the influence of built environment variables on trip demand as well as on the ratio of demand to supply at bike stations in China. Faghih-Imani et al. [15] investigated factors affecting bicycle share demand at the station level using real-time ridership data. The results showed that stations close to major roads had lower trip activities compared to stations that were situated around minor roads and bicycle lanes. A number of land use and built environment variables, temporal characteristics, and weather variables such as temperature were investigated. Maurer [20] used a pairwise suitability analysis to understand the effects of variables such as job density, household income, and alternative commuters on public bicycle share ridership to propose the locations of bicycle stations in

Sacramento, California. Gebhart and Noland [21] used real-time ridership data for Capital Bikeshare in Washington D.C. to investigate the impact of weather variables and proximity of bike share stations to metro stations on ridership levels. Buck and Buehler [22] investigated the influence of bicycle infrastructure, population density, land use mix around stations, and the number of households without a car using bicycle share systems using ridership data from Capital Bikeshare. Wang et al. [23] evaluated the effect of sociodemographic, land use, built environment, and transportation infrastructure variables on bicycle share ridership. Rixey [24] explored the influence of sociodemographic characteristics such as education, income, and employment and population density on monthly ridership data from three states of the USA.

2.3. Spatial and Temporal Patterns of Bike Use. These studies explore the spatial and temporal patterns of bike use over the time of day, using data mining and visualization techniques. Clustering is frequently used to identify mobility patterns in BSS usage by partitioning the stations into different clusters having a similar usage. Wong and Cheng [25] presented the insights of imbalanced public bicycle distributions through the analysis of spatiotemporal activity patterns of bike stations. The clustering algorithm was used to analyze how station activity patterns were geographically distributed based on their usage patterns. They also explored how these activity patterns relate to underlying cultural and spatial characteristics of Taipei City in China. Temporal and spatiotemporal patterns among bike stations of Barcelona bike sharing system were explored by Froehlich et al. [26]. Numerous research studies also used a hierarchical clustering method to generate clusters and investigate usage patterns geographically distributed in the city to understand the impact of the inhomogeneity of the city on the long-run activity of stations [27–29]. Brien et al. [30] proposed a classification of bike shares based on the geographical footprint and diurnal, day-of-week, and spatial variations in occupancy rates. Etienne and Latifa [31] presented an automatic algorithm based on a new statistical model to automatically cluster PBS stations according to their usage profile. Zhou [32] investigated the spatiotemporal biking pattern in Chicago by analyzing massive BSS data from July to December in 2013 and 2014. Bike flow similarity graph was constructed with a fast greedy algorithm to detect spatial communities of biking flows.

Scholars have achieved rich results in measuring the indicators of the bicycle system and the factors affecting cycling. The methods of research mainly involve regression models. The knowledge gained most comes from the dock PBS except a few studies [33–35].

3. Methodology

3.1. Indicators of Dockless PBS. There are many indicators to measure the PBS, including the number of bicycle use, arrival rate, and departure rate. This study focused on the number of available bicycles and their changes, so two indicators were used.

3.1.1. Average Available Bike Number. Unlike the dock PBS, the maximum available bike number is fixed and determined by the number of docks of station. For the dockless PBS, the maximum available bike is not subject to parking restrictions. It is related to the initial bike quantity status of deployment by the system and varies as the bike flows. We proposed the average available bike number to explore the dockless PBS. It represents the number of bicycles available in a bike service area. This metric is used to measure the bicycle resource. The number of bikes available per hour (Abn_i) can be calculated by equation (1).

$$abn_i = \frac{\sum_{d=1}^{5} abn_d^i}{5}, \tag{1}$$

$$abn\_\mathrm{DAY} = \frac{\sum_{i=1}^{24} Abn_i}{24}, \tag{2}$$

where i represents the ith hour in a day; d represents the dth workday of a week; Abn_d^i is the number of available bicycles in the ith hour of the dth day; Abn_i is the average number of available bicycles in hour i in work day; and abn_DAY in equation (2) is the average available vehicle for bike service area throughout the day.

3.1.2. Coefficient of Available Bike Number Variation. The coefficient of variation is used to compare the degree of dispersion of the two sets of data, which can eliminate the influence of measurement scale and dimension. In this study, the coefficient of variation was used to compare the changes in available bicycles in 24 hours of a day among bicycle service areas. The calculation formula is shown in the following equation:

$$cv = \frac{\mathrm{s.d}\left(abn_i\right)}{abn\_\mathrm{DAY}} \times 100, \tag{3}$$

where $\mathrm{s.d}\left(abn_i\right)$ is the standard deviation of available bicycles number in 24 hours in a service area. Obviously, cv is affected by the two statistics of mean and standard deviation of available bike number. This metric is used to measure the variation in bicycle usage with average available bike number.

3.2. Clustering and Classification in Machine Learning Approach

3.2.1. Clustering Algorithm. Clustering is an unsupervised learning algorithm of classifying and organizing members in datasets which are similar [36].

(1) k-Means Clustering Algorithm. Given a set of data, the k-means algorithm divides the data into k clusters repeatedly according to a distance function. The algorithm operates on a set of d dimensional vectors, $D = \{x_i | i = 1, \ldots, N\}$, where $x_i \in d$ denotes ith data point. The algorithm is initialized by picking k points in d as the initial k cluster representatives or "centroids." Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as

the solution of clustering a small subset of the data, or perturbing the global mean of the data k times. Then, the algorithm iterates between two steps till convergence. About the value of k, we can choose from reasonable guessed or predefined number, but it is better to know whether k clusters is better or worse than $k-1$ or $k+1$ clusters. The method of With he Sum of Square (WSS) is often used to get reasonable K value. WSS is the sum of the square of the distance between all points and their nearest centroid point. The calculation is shown in equation (4). p_i represents the point i, and q^i represents the nearest centroid point to i; if all data points are relatively close to their respective centers, then the WSS is relatively small. If $K+1$ clusters do not significantly reduce the WSS value of K clusters, then the classification is of little significance.

$$\mathrm{WSS} = \sum_{i=1}^{N} d\left(p_i, q^i\right)^2. \tag{4}$$

(2) Mean Shift Clustering Algorithm. Mean shift clustering is a general nonparametric cluster finding procedure introduced by Fukunaga and Hostetler [37], and it does not depend on any explicit assumptions on the shape of the point distribution, the number of clusters, or any form of random initialization. Mean shift treats the clustering problem by supposing that all points given represent samples from some underlying probability density function, with regions of high sample density corresponding to the local maxima of this distribution. To find these local maxima, the algorithm works by allowing the points to attract each other, via what might be considered a short-ranged "gravitational" force. Allowing the points to gravitate towards areas of higher density, one can show that they will eventually coalesce at a series of points, close to the local maxima of the distribution. Those data points that converge to the same local maxima are considered to be members of the same cluster. For a mathematical details, see Comaniciu and Meer [38]. In the next sections, we illustrate application of the algorithm to a couple of problems using the python package SkLearn which contains a mean shift implementation.

3.2.2. Classification Algorithm. Classification is a kind of supervised learning algorithm of training a classifier in a group of samples that already know the class label so that it can classify an unknown sample. In the field of machine learning, there are hundreds of classifiers to solve real-world classification problems [39], and in this research, five commonly used classification algorithms are selected: random forest classifier, K-nearest neighbor classifier, logistic regression, support vector machine, and artificial neural network. The five algorithms used in this study are based on the Python platform Scikit-learn package free from https://scikit-learn.org. The parameters of each algorithm have been adjusted to ensure the optimal performance of the algorithm. In the analysis of shared bicycles, the accuracy and robustness of these five classification algorithms will be compared.

(1) Random Forest Classifier. Random forest classifier (RFC) is the most widely used supervised machine learning algorithm. It is very powerful and usually gives good results without the need to repeatedly adjust the parameters. The basic unit of random forests is the decision tree. A random forest is a classifier that contains multiple decision trees, and the category of its output is determined by the mode of the category of the individual tree output [40]. For an input sample, N trees will have N classification results. The random forest integrates all the classification voting results and specifies the category with the highest number of votes as output. It has several advantages: it enables to handle thousands of input variables without variable deletion and gives estimates of what variables are important in the classification.

(2) K-Nearest Neighbor Classifier. K-nearest neighbor (KNN) is a method of measuring the distance between different feature values for classification. Given a training set D and a test object z, the test object is a vector composed of attribute values and an unknown category label. The algorithm needs to calculate the distance (or similarity) between z and each training object. In this way, the list of nearest neighbors can be determined. Then, assign the category with the dominant number of instances in the nearest neighbor to z. The advantage is that it is easy to understand, and good performance can be obtained without excessive adjustment. The disadvantage is that the prediction speed is slow and the dataset with many characteristics cannot be processed. It is vulnerable to data imbalance. And the interpretability of the output is not strong.

(3) Logistic Regression. Logistic regression (LR) is essentially a linear classifier, which refers to the establishment of a regression formula on the classification boundary line based on the existing data to classify. The calculation cost of this method is not high, and it is easy to understand and implement. The fitted parameters can clearly see the impact of each feature on the result. And most of the time is used for training, and classification is fast after training is completed, but it is easy to underfit and the classification accuracy is not high. The main reason is that LR is linear fitting, but in reality, many things do not satisfy linearity.

(4) Support Vector Machine. Support vector machine (SVM) maps the data to a multidimensional space in the form of points, thereby converting the nonlinear separable problem in the original sample space into a linear separable problem in the feature space so that the optimal hyperplane for classification can be found. Then, classify the set according to the hyperplane. SVM can make good predictions on data outside the training set and has a low generalization error rate, low computational overhead, and easy-to-interpret results, but it is too sensitive to parameter adjustments and kernel function parameters.

(5) Artificial Neural Network. Artificial neural network (ANN) is an information processing system based on imitating the structure and function of the brain's neural

network. The ANN algorithm is a set of continuous input/output units, where each connection is associated with a weight. In the learning stage, by adjusting the weights of the neural network, the correct class label of the sample to be learned can be predicted. The advantages of the ANN algorithm are high classification accuracy and strong distributed parallel processing capabilities. Artificial neural networks have strong robustness and fault tolerance for datasets containing a large amount of noisy data, but the learning process cannot be observed, and the output results are difficult to interpret, which will affect the reliability and acceptability of the results. It also requires a large number of parameters, such as network topology, initial values of weights, and thresholds.

4. Study Area

4.1. OFO Dockless PBS in Shenzhen. This paper focuses on China's fastest urbanizing city, Shenzhen, to lay a foundation for empirical analysis of the intensity of usage of the OFO dockless bike sharing system. It provides a unique case study as it is one of the largest bike share programs located in a metropolis. OFO bicycle sharing system was launched in Shenzhen in December 2016 with more than 2200,00 bicycles. We scanned the working status of these bicycles every 15 minutes in one week of September 2017. There are about 57.6 million bicycle status records in a day. For a bicycle ID, we first judge whether the bicycle is used by comparing whether its position has changed. If changed, we saved the time and position of the bicycle. Then, according to the average travel speed and travel distance of the bicycle, the abnormal bicycle use record is rejected. Figure 1 demonstrates the bike service area in Shenzhen.

Figure 2 shows the trip summary of shared bicycles in 24 hours in a workday. There are two distinct peaks in shared bicycle use in a workday. The morning peak is between 08:00-09:00, and the evening peak is between 18:00–21:00. It is reasonable to assume that bikes are used for commuting. In the morning peak, the trip number of bicycles exceeded 50,000. The trips in evening peak were slightly lower than the early peak, but still more than 40,000. During the period of 01:00–06:00, bicycle usage is stable and the lowest with about 5,000 trips per hour. At noon period from 12:00 to 15:00, the bicycle use is about 20,000 per hour. The amount of bicycle use dropped from 40,000 to 10,000 per hour at the night period which is from 22:00 to 24:00.

4.2. Influencing Factor of Bike Use. In the previous studies, factors influencing public bike usage are grouped into four categories: transportation, land-use/build environment, population, and meteorological data. The weather variables are not considered in our study. A total 25 factors were selected including 6 categories of variables: population, point of interest (POI), road network, public transportation, distance, and building function. The detailed factors are listed in Table 1.

5. Results and Discussion

5.1. The Identification of Bicycle Gathering Area. We used the mean shift clustering method to identify the clustering area of bicycles based on the position of the bicycle at 09:00. In the choice of bandwidth, we considered two bandwidths: 300 meters and 500 meters, because the area identified by these two bandwidths is approximately equal to the grid area size of $500\,\text{m} * 500\,\text{m}$ and $1000\text{m} * 1000\,\text{m}$. The minimum number of bicycles included in each category is set to 100. When the bandwidth is 300 meters, a total of 492 bicycle gathering areas are obtained. The 492 bicycle gathering areas contain a total of 140,000 bicycles, accounting for 63.6% of all bicycles. When the bandwidth is 500 m, a total of 270 gathering areas are obtained, including 140,000 bicycles. Considering that the bicycles contained in the 492 clusters are more compact, we finally selected 492 clusters as the analysis objects.

Figure 3 shows the OFO bicycle gathering area identified by mean shift clustering. In Figure 3, each cluster has a center point and the buffer analysis was proposed to obtain the range of the bicycle gathering area. The buffer is a kind of influence range or service scope of the geospatial target, which refers to the polygons of a certain width which are automatically established around the point, the line, and the surface entity. 300 meters of buffers were established by ArcGIS based on all cluster center points, thus to calculate the indicators of dockless PBS and influencing factor of bicycle gathering area.

5.2. Performance of Five Classification Algorithms. After calculating the available bike number and coefficient of available bike number variation of bicycle cluster area, the k-means algorithm was executed to group these areas. Figure 4 shows the WSS curve, and we made WSS values from 2 clusters to 19. When k is increased from 2 to 8, WSS decreases significantly. When $k > 8$, the improvement of WSS is very linear so the cluster centers have similar characteristics. The larger the k means the more the classifications of bicycle cluster area which is likely to impact on the accuracy of the classification algorithm. It is necessary to find an optimal k value to balance between the accurate cluster and accurate classification prediction. This research adopts an experimental strategy to select k from 3 to 8 and then uses five classification algorithms to compare the prediction accuracy. The bicycle gathering areas in the same cluster are marked with the same label using k-means clustering. Five classification algorithms were compared in the accuracy of distinguishing the type of bicycle gathering areas using 25 impact factors. The experimental process is divided into two stages including training and application. At the stage of training, the 492 gathering areas are randomly divided into two parts. The first part containing 75% of areas is used for training data, and the second part as the test data is used to verify the accuracy. Figure 5 shows the accuracy of the five classification algorithms in the training set and the test set when k takes different values.

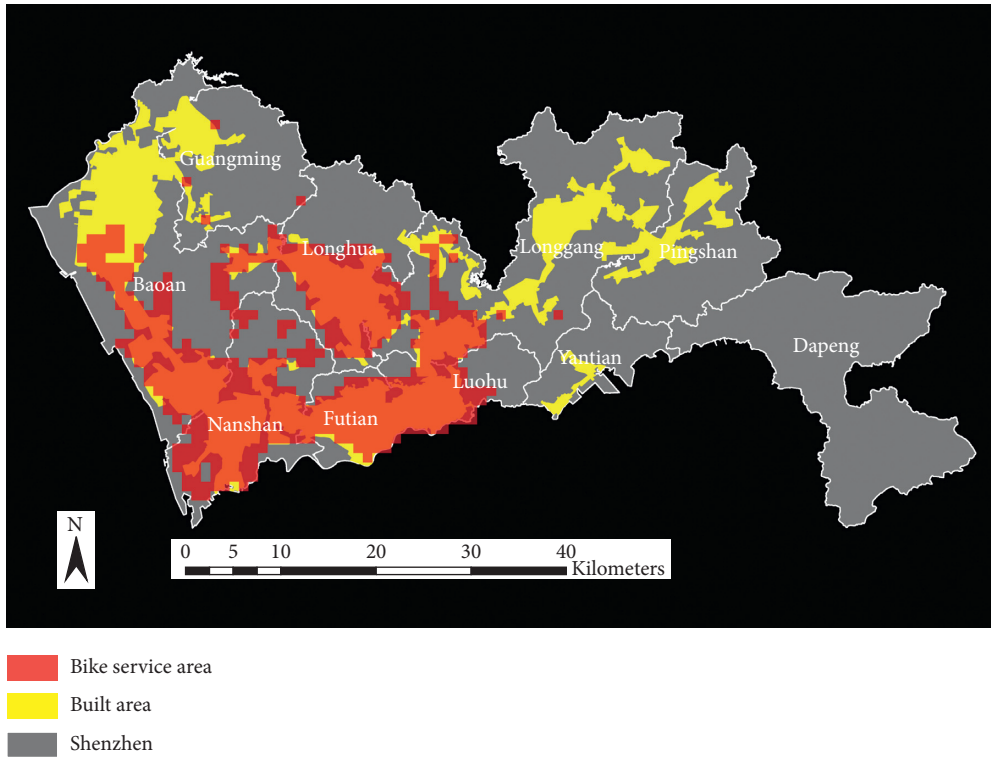

Figure 1: OFO bike service area in Shenzhen.

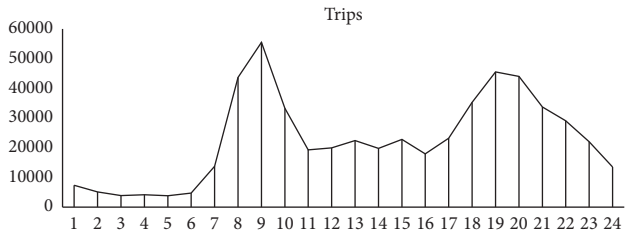

Figure 2: The number of sharing bicycle trips in 24 hours in a workday.

In the training set, the performance differences of the five algorithms are obvious. For different K values, the RFC always maintains the highest accuracy rate, which is higher than 90%. The ANN also has a high accuracy rate. When K is 3-4, the accuracy rate is above 90%, and when the k value is 5–8, the accuracy rate drops to above 80%. KNN algorithm performance is in the middle of the five algorithms. When the K value is 3-4, the accuracy rate is above 70%, and when the K value is 5–8, the accuracy rate drops to above 60%. As the k value increases, the accuracy of SVM drops from 63% to 48%. The worst-performing algorithm is the LR algorithm. As the k value increases, the accuracy rate drops from 58% to 37%. In addition, in the trend that the accuracy rate changes with the value of k, the accuracy rate of the RFC algorithm fluctuates little, and the other algorithms have the highest accuracy rate when the value of K is small. As the value of k increases, the accuracy rate decreases, and When $k = 8$ with ANN, the accuracy rate has increased. In the test set, the accuracy of the five algorithms is lower than that of the training set. The most accurate performance is still the RFC algorithm. When $k = 4$, RFC has the highest accuracy rate of 76.97%, which is the only case in the test set where the accuracy rate exceeds 75%. When $k = 3$, its accuracy rate is 73%. For ANN, the highest accuracy rate is 71% when $k = 4$. The accuracy of KNN in the training set is better than that of SVM and LR, but its performance in the test set is not much different from SVM and LR. The accuracy of these three algorithms is low. In addition, RFC, ANN, and LR all have the highest accuracy when $k = 4$.

Comprehensive comparison of the performance of the five algorithms in the training set and the test set shows that RFC and ANN have better performance in the prediction of the type of bicycle clusters. The accuracy of ANN, SVM, and LR in the training set is quite different, but the difference in the test set is not obvious. Basically, when the k value is greater than 4, as the K value increases, the accuracy of the five classification models has a downward trend. When $K = 4$, RFC and ANN have the highest test accuracy. We choose $k = 4$, which means that the bicycle aggregation is divided into 4 types for further analysis.

5.3. The Analysis of Bicycle Gathering Area

5.3.1. The Clustering of Bicycle Gathering Area. Table 2 shows the description of the cluster center when $k = 4$. The table mainly lists four indicators, which are the standard and original values of Abn_DAY, and the standard and original values of the coefficient of variation (in k-means clustering, standard values were used).The four cluster centers have obvious characteristics. We first divide

Table 1: The influencing factors of bike use.

Factor type	Factor	Unit	Calculation
Population	Population	Number	Count the number of residents in the service area
POI	Restaurant	Number	Count the number of POIs of the corresponding category in the service area
	Company	Number	
	Small store	Number	
	Car park	Number	
Road network	Length of main road	m	Calculate the total length of the corresponding road level in the service area
	Length of secondary road	m	
	Length of branch road	m	
Public transportation	Bus stop	Number	Calculate the number of bus stops in the service area
	Distance to subway	m	Calculate the distance from the center of the service area to the nearest subway station
Distance	Distance to university	m	Calculate the distance from the center of the service area to the closest corresponding place
	Distance to government	m	
	Distance to supermarket	m	
	Distance to hub	m	
	Distance to square	m	
	Distance to park	m	
	Distance to school	m	
	Distance to hospital	m	
Building function	Office building	m^2	Calculate the total floor area of the corresponding building in the service area
	Industrial building	m^2	
	Public building	m^2	
	Commercial building	m^2	
	Residential building	m^2	
	Urban village building	m^2	
	Warehouse	m^2	
	Building number	Number	
	Cover ratio	%	The ratio of the projected area of all buildings to the area of the service area

the four clusters into inefficient and efficient groups according to the value of *cv*. A high *cv* indicates that the number of available bicycles in the gathering area is more appropriate, and the use of regional bicycles is efficient. A low cv indicates that the number of available bicycles in the gathering area is large, which does not match the number of active bicycles, and the use of bicycles is inefficient. We call clusters with z_*cv* lower than 0 as an inefficient group and clusters with z_*cv* greater than 0 as an efficient group. Then, each group is divided into two subtypes according to *Abn*_DAY and *cv*.

(i) Cluster A: $z\_Abn\_\mathrm{DAY} > 1$, $z\_cv < 0$: this cluster can be called high efficiency mode. *Abn*_DAY in this group is reaching 416, but the average daily change of vehicles is very few, with an average of only 51.There are excessive bicycles deployed or stayed in the area, and the activity of more than 300 bicycles is not high.

(ii) Cluster B: $z\_Abn\_\mathrm{DAY} < 0$, $z\_cv < 0$: we call it a normal inefficient mode. Its $z\_cv < 0$ is as same as Cluster A, but $z\_Abn\_\mathrm{DAY} < 0$ compared with A indicating that the number of bicycles in the cluster area in this group is less than A. There are about 185 bicycles with an average of 17 bikes daily used, and more than 150 are not very active.

(iii) Cluster C: $z\_Abn\_\mathrm{DAY} < 0$, $z\_cv > 2$: this cluster has the highest *cv* value among the four clusters, indicating that the number of available bicycles in this group matches the demand for bicycles, and there are not too many idle bicycles available. The average number of available bicycles in this group is 213, and the average daily change is 117. More than half of the bicycles are used, so it is called high efficient mode.

(iv) Cluster D: $z\_Abn\_\mathrm{DAY} < 0$, $z\_cv > 0$: this cluster is similar to C, except that the *z_cv* value is lower than that of the C class but exceeds the average. In this group, the average number of available bicycles is almost similar to C, but *Abn*_DAY is about half of that of C, and its average *cv* is 2-3 times that of groups A and B. The use efficiency of bicycles is higher than that of A and B but lower than C, so it is called normal effective mode.

In general, high inefficient mode of cluster A has the max average number of available bikes and high efficiency mode of cluster C has the largest average coefficient of variation. The difference between A and B is in the average number of available bikes, and the difference between B, C, and D is in the coefficient of variation. From the number of the clusters, cluster A and cluster B together account for about 73%, so

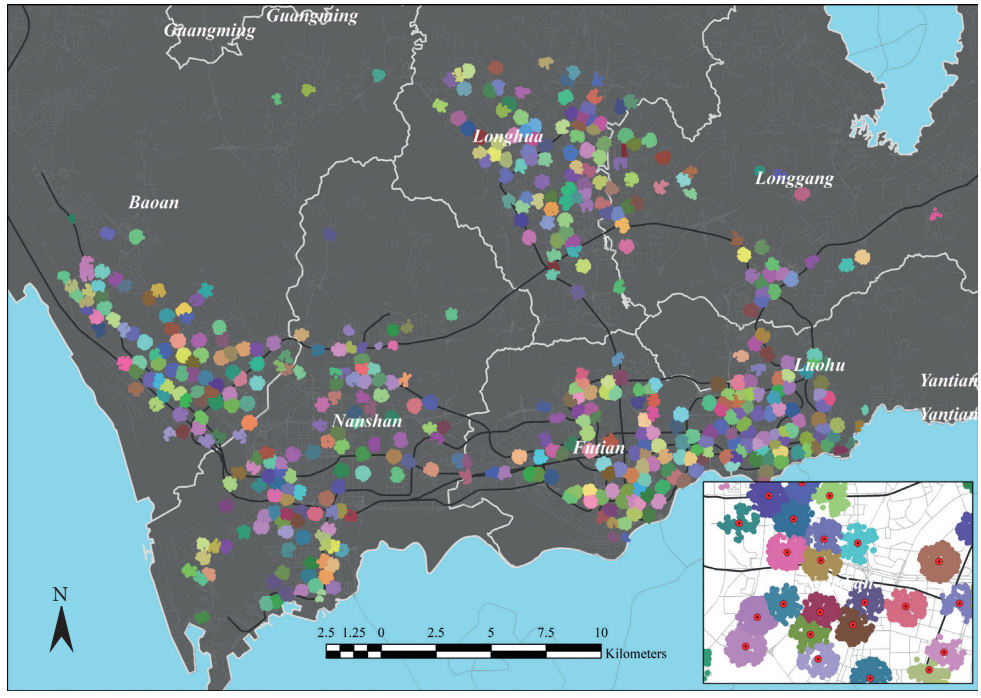

Figure 3: The OFO bicycle gathering area identified by mean shift clustering.

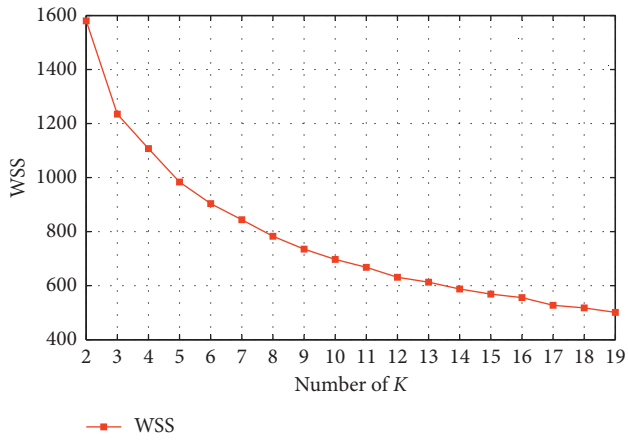

Figure 4: The WSS curve in determining the number of clusters.

the main mode is inefficient mode. These two groups have gathered a total of 110,000 bicycles which used low efficiency. The efficient mode area accounts for about 27% of all regions, of which the high efficient mode area accounts for 10%, with 10,000 bicycles, and the normal efficient mode accounts for 17%, totaling 17,705 bicycles. A total of 27,000 bicycles are gathered in these two groups, and the bicycle use efficiency in these areas is higher.

5.3.2. The Impact of Clustering of Bicycle Gathering Area. Figure 6 shows the spatial distribution of four clusters. The high inefficient mode is mainly distributed in the central area of Shenzhen (Futian and Luohu) and Baoan and shows the characteristics of spatial clustering. Normal inefficient areas are distributed on the periphery of urban built-up areas, and efficient mode areas are scattered between normal inefficient and high inefficient areas. It is worth noting that in Futian and Luohu districts where the subway network density is high, the main distribution mode is inefficient. In order to better understand the influencing factors that affect the types of bicycle service clusters, we analyzed the importance of the factors that determine the types of bicycle service clusters based on the RFC model with the highest accuracy.

The importance indicates how important the variables are in the RFC model. The sum of the importance of all variables is 1. Figure 7 shows the importance of 25 factors in the RFC model in ascending order when $K = 4$ and the average importance is 0.04 (1/25). The importance of population and buildings number is significantly higher than other factors which are key factors. The least important factors are the length of the branch road, the number of bus stops, and the area of public buildings. The importance of these three variables does not exceed 0.02 lower than the average. The importance of the main road length and the restaurant number ranked third and fourth, indicating that they are important reference variables for identifying bike gathering area type. The importance of resident building area, building coverage, distance to universities, and small shops is slightly larger than average. The distance to the subway station ranks only ninth in importance, which is about the same as the commercial building area and the company number. The sum of the importance of the top 10 variables accounted for 54%. Existing studies have shown that the area around subway stations is the most active area

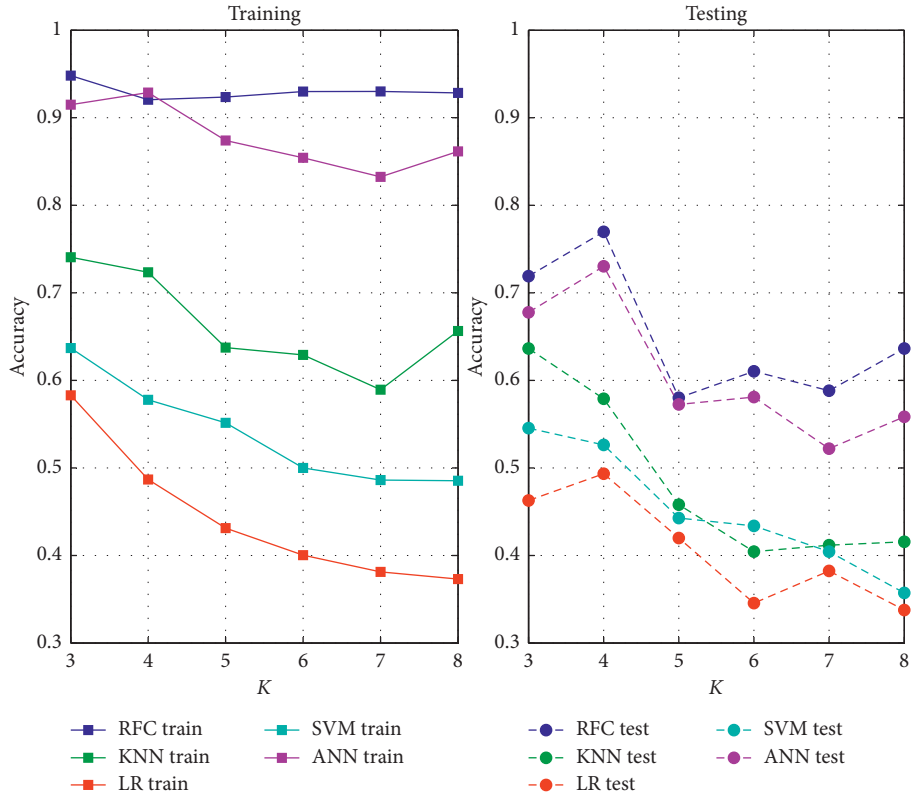

Figure 5: The accuracy of the five classification algorithms.

Table 2: The description of four clusters.

Cluster	A	B	C	D
Description	High inefficient	Normal inefficient	High efficient	Normal efficient
*z_ Abn_*DAY	1.03	−0.73	−0.52	−0.55
z_ cv	−0.39	−0.59	2.35	0.75
Abn_DAY	416	185	213	209
cv	0.12	0.09	0.55	0.30
Abn_DAY ∗ *cv*	51	17	117	63
Total available bike number	78588	31574	9986	17705
Cluster number	189	171	47	85

for bicycle activities, but our research has found that the distance from the subway station is not the most important factor in judging the activity type of bicycle gathering areas. The population, the number of buildings, the length of the main road, and restaurants are four important variables for judging the active types of bicycle clusters. In addition, among the 25 influencing factors in Figure 7, except for the variables with higher and lower importance, the importance of most of the variables in the middle is more evenly distributed, indicating that the active types of bicycle clusters have more and more complex influencing factors.

Figure 8 shows a comparison of the average values of the standard values of 25 factors. The color of the heat map clearly shows that there are obvious differences in the values of 25 variables between the four groups. We found that extreme values of variables tend to appear in groups A and C. Group A is significantly higher than the other three groups in the seven variables of population, number of buildings, length of main roads, number of restaurants, building coverage, number of parking lots, and number of bus stops. The number of population buildings and the length of secondary roads in group C are significantly lower than the other three groups, and the distance to school, industrial building area, office area, and distance to the park are significantly higher than the other three groups. Although the variables in groups B and D rarely have maximum or minimum values, the characteristics of the variables between them are quite different. The classification is based on average available number and the coefficient of variation, but the 25 variables between the classes

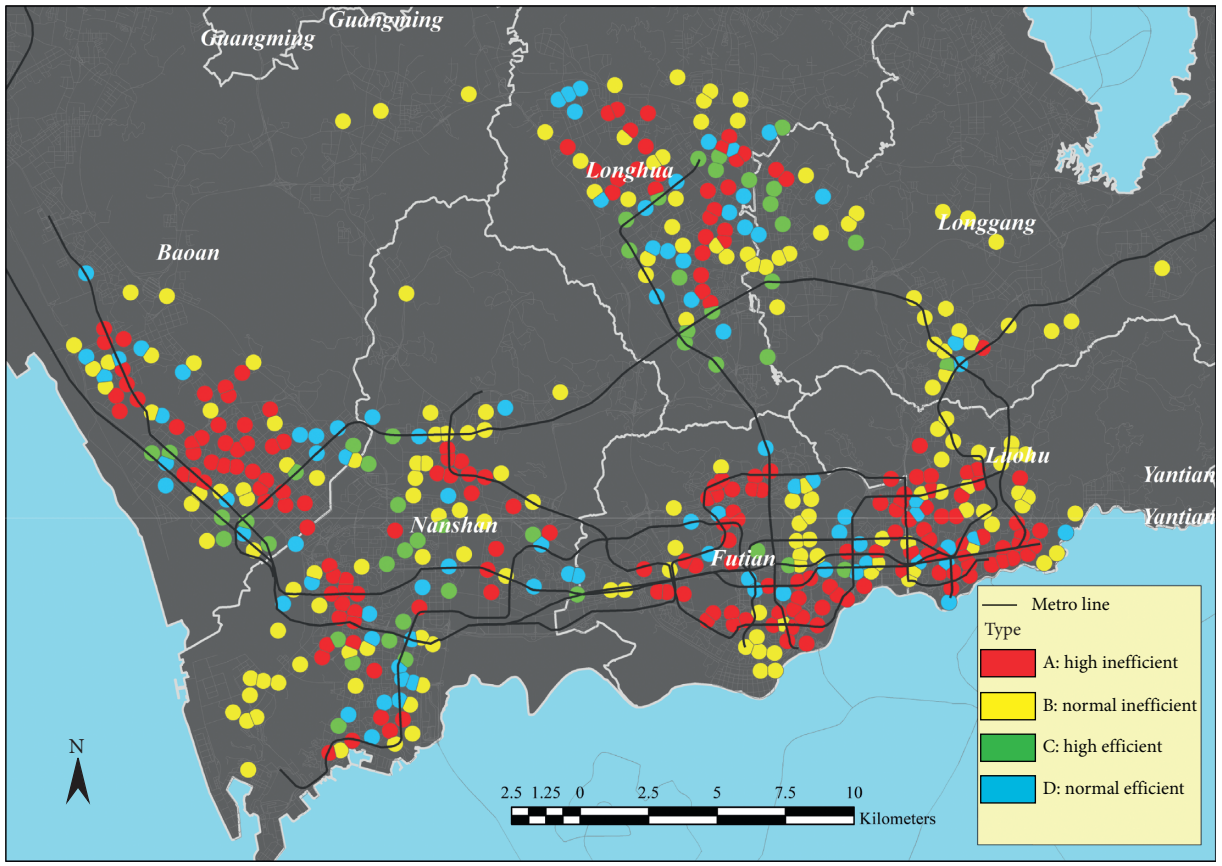

Figure 6: The distribution of four clusters.

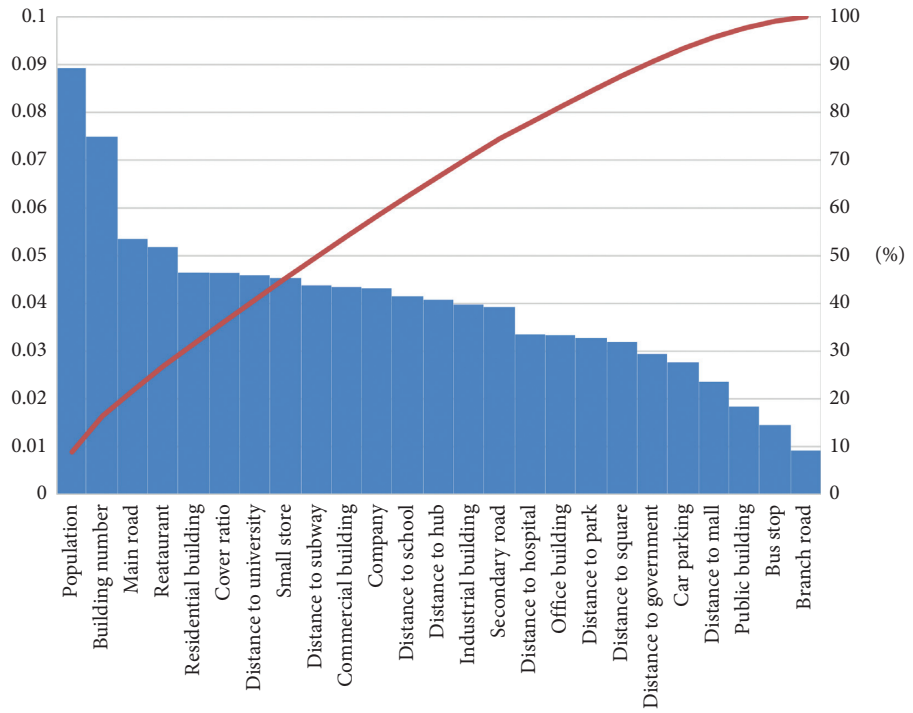

Figure 7: The importance of factor in random forest classifier model.

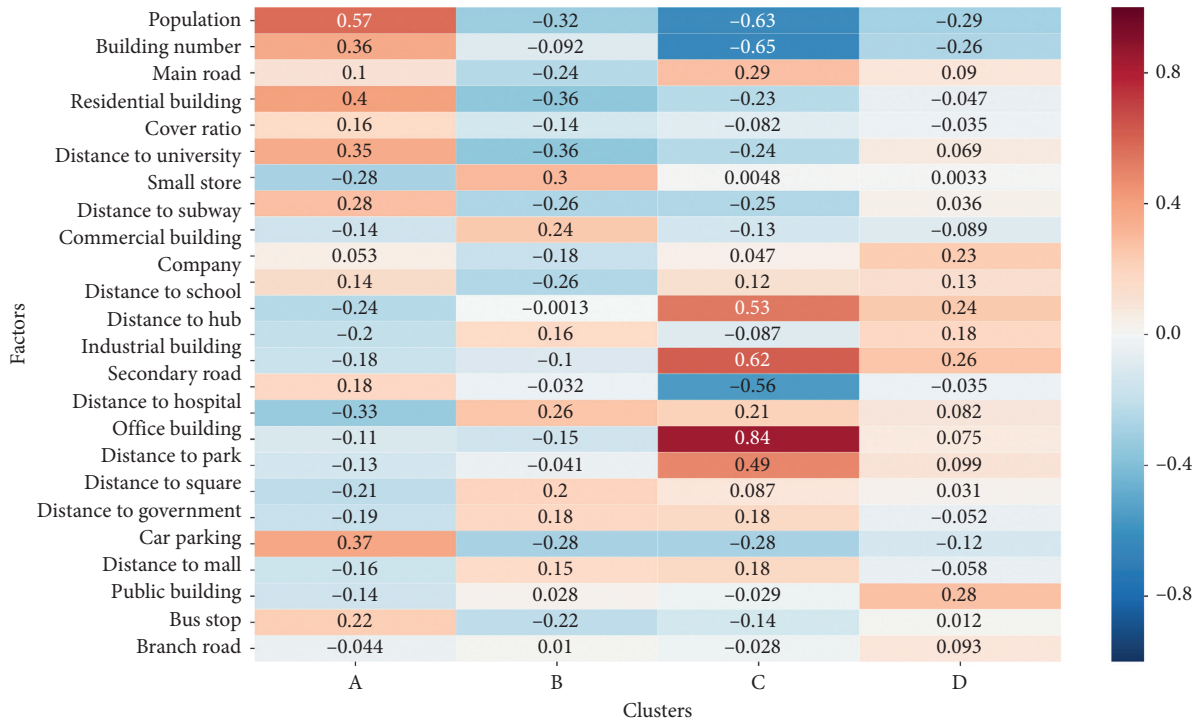

Figure 8: The heatmap of factor in four clusters.

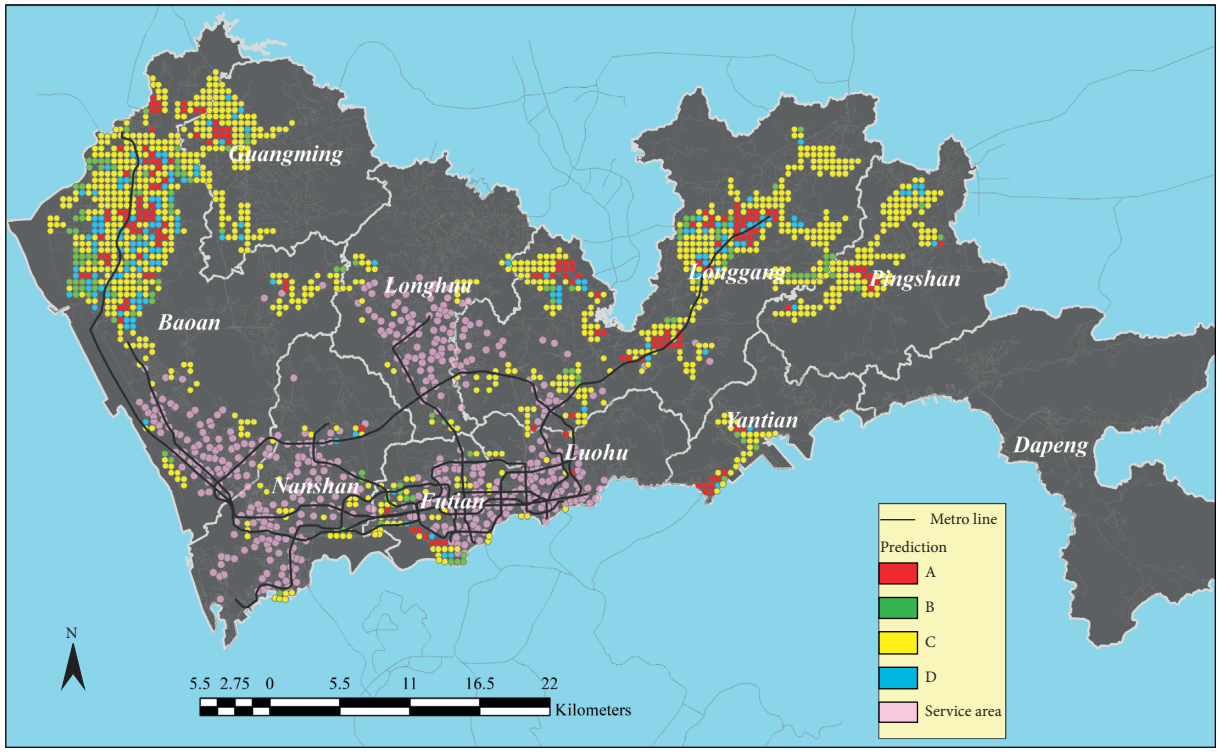

Figure 9: The predicted results of activity patterns in the new service area.

have obvious differences in value, indicating that the activity types of bicycles are related to these factors. The values of these variables can be used to judge the activity types of bicycles.

5.3.3. Guiding Public Bicycle Planning and Management. Let us assume a scenario: to provide a dockless public bicycle service in a new area in Shenzhen. Considering that the influencing variables are easy to obtain, we apply the RFC to

predict the activity pattern of the new service area. The built area of Shenzhen which does not provide OFO service is divided into 1459 grids, and the influence factors are calculated. Figure 9 shows the predicted results of the RFC model for activity patterns. Note that the prediction assumes that the existing bike operational and deployment strategies of OFO remain unchanged. Yellow grids belonging to cluster B have the largest number about 1025. There are 146 blue grids which are cluster D. The remaining area is 162 red grids and 126 green grids. Most girds are normal inefficient types. Table 2 and the service area type can provide information to public bicycle quantity management. We can get theoretically the minimum number of bicycles needed according to the grid's cluster and the total number of bicycles required. It is possible to optimize the number of available bicycles according to bike activity of the grid, reduce operating costs, and improve utilization efficiency.

6. Conclusions

This research practiced the introduction of a machine learning approach to quantity management using OFO bike operation data in Shenzhen. The contributions are mainly reflected in the following three aspects. First, we proposed a method for identifying the cluster area of dockless shared bicycles, which can accurately calculate the impact factors of shared bicycle systems. Second, different from previous research perspectives, this research discusses the performance and optimization possibilities of the shared bicycle system from the number of available bicycles in the gathering area and its changes. At last, this work shows the applicability and operability of machine learning methods in solving urban planning and management problems, which is inspiring for people with urban management background to use computational intelligence.

The bicycle gathering area type's recognition, prediction, and application in this study are meaningful for the sustainable development of shared bicycles. (1) There were 492 OFO bicycle gathering areas containing more than 140,000 bicycles, accounting for 63.6% of all bicycles in Shenzhen. (2) More type number of bicycle gathering area will affect the accuracy of the classification algorithm. The random forest classification had the best performance in identifying bicycle gathering area with an accuracy of more than 75%. (3) Shenzhen OFO dockless public bike gathering areas can be divided into four types: high inefficient, normal inefficient, high efficient, and normal efficient. The main area types are high inefficient and normal inefficient which gathered about 110,000 bicycles with low usage.(4) There were obvious differences in the characteristics of impact factors in four types of bicycle gathering areas. It is feasible to use these factors to predict area type to optimize the number of available bicycles, reduce operating costs, and improve utilization efficiency. So, the knowledge from the existing dockless bicycle operation data can be used to guide public bicycle planning and management. The potential activity patterns and the minimum number of bikes in new service areas can be obtained in advance. Operating companies can make optimization strategies based on this information.

Our study also has some limitations. First, due to the limitations of data acquisition, the working day operational data used only contain one week, so the results of the analysis may be biased. The data we analyzed did not include data for nonworking days. Weekend public bicycle usage patterns may differ from weekday. Second, the modes analyzed in this study rely heavily on operational data and may not be applicable elsewhere. When the strategy of bicycle operation changes or the number of bicycles is reoptimized, the activity modes will be affected. After a period of operation, according to the indicators and models of this study, a new mode of activity will be formed. Last, this paper focuses on the number of available bicycles, and the activity indicators of dockless PBS need to be further explored.

Data Availability

The bike data, population distribution data, and POI data including land use and built environment were supplied by the authors of this article. They are freely available. Requests for access to these data should be made to Qingfeng Zhou, zhouqingfeng@hit.edu.cn.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Financial support from the National Natural Science Foundation of China (grant no. 41771169) is acknowledged.

References

[1] P. Demaio, "Bike-sharing: history, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, pp. 41–56, 2009.

[2] China Academy of Information and Communications Technology, *China Shared Bicycle Industry Development Report*, China Academy of Information and Communications Technology, Beijing, China, 2018.

[3] Y. Zhou, B. P. L. Lau, Z. Koh, C. Yuen, and B. K. K. Ng, "Understanding crowd behaviors in a social event by passive WiFi sensing and data mining," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4442–4454, 2020.

[4] K. Li, C. Yuen, S. S. Kanhere et al., "An experimental study for tracking crowd in smart cities," *IEEE Systems Journal*, vol. 13, no. 3, pp. 2966–2977, 2019.

[5] Q. Chen, W. Wang, F. Wu et al., "A survey on an emerging area: deep learning for smart city data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 5, pp. 392–410, 2019.

[6] Y. Zhou, B. P. L. Lau, C. Yuen, B. Tuncer, and E. Wilhelm, "Understanding urban human mobility through crowdsensed data," *IEEE Communications Magazine*, vol. 56, no. 11, pp. 52–59, 2018.

[7] L. P. L. Billy, N. Wijerathne, B. K. K. Ng et al., "Sensor fusion for public space utilization monitoring in a smart city," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 473–481, 2017.

[8] X. Wang, C. Yuen, N. U. Hassan et al., "Electric vehicle charging station placement for urban public bus systems,"

IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 1, pp. 128–139, 2017.

[9] A. Faghih-Imani and N. Eluru, "Analysing bicycle-sharing system user destination choice preferences: chicago's Divvy system," *Journal of Transport Geography*, vol. 44, pp. 53–64, 2015.

[10] C. Fricker and N. Gast, "Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity," *EURO Journal on Transportation and Logistics*, vol. 5, no. 3, pp. 261–291, 2016.

[11] P.-S. You, P.-J. Lee, and Y.-C. Hsieh, "An artificial intelligent approach to the bicycle repositioning problems," *Engineering Computations*, vol. 34, no. 1, pp. 145–163, 2017.

[12] E. O'Mahony and D. B. Shmoys, "Data analysis and optimization for (citi) bike sharing," in *Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence*, AAAI Press, Austin, TX, USA, January 2015.

[13] Y. Chen, Y. Li, H. Hu, J. Zhang, D. Gu, and P. Xu, "Computational intelligence approaches to robotics, automation, and control," *Mathematical Problems in Engineering*, vol. 2015, Article ID 620275, 1 page, 2015.

[14] Áo Lozano, J. De Paz, G. Villarrubia González, D. Iglesia, and J. Bajo, "Multi-agent system for demand prediction and trip visualization in bike sharing systems," *Applied Sciences*, vol. 8, no. 1, p. 67, 2018.

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

[16] G. R. Krykewycz, C. M. Puchalsky, J. Rocks, B. Bonnette, and F. Jaskiewicz, "Defining a primary market and estimating demand for major bicycle-sharing program in philadelphia, Pennsylvania," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2143, no. 1, p. 117, 2010.

[17] W. El-Assi, M. Salah Mahmoud, and K. Nurul Habib, "Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto," *Transportation*, vol. 44, no. 3, pp. 589–613, 2017.

[18] R. C. Hampshire and L. Marla, *An Analysis of Bike Sharing Usage: Explaining Trip Generation and Attraction from Observed Demand*, Transportation Research Board, Washington, DC, USA, 2012.

[19] Y. Zhang, T. Thomas, M. Brussel, and M. van Maarseveen, "Exploring the impact of built environment factors on the use of public bikes at bike stations: case study in Zhongshan, China," *Journal of Transport Geography*, vol. 58, pp. 59–70, 2017.

[20] L. K. Maurer, *Feasibility Study for a Bicycle Sharing Program in Sacramento, California*, Transportation Research Board, Washington, DC, USA, 2011.

[21] K. Gebhart and R. B. Noland, "The impact of weather conditions on bikeshare trips in Washington, DC," *Transportation*, vol. 41, no. 6, pp. 1205–1225, 2014.

[22] D. Buck and R. Buehler, *Bike Lanes and Other Determinants of Capital Bikeshare Trips*, Transportation Research Board, Washington, DC, USA, 2012.

[23] X. L. G. S. Wang, "Modeling bike share station activity: effects of nearby businesses and jobs on trips to and from stations," *Journal of Urban Planning & Development*, vol. 142, no. 1, 2012.

[24] R. A. Rixey, "Station-level forecasting of bikesharing ridership," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2387, no. 1, pp. 46–55, 2013.

[25] J. T. Wong and C. Y. Cheng, "Exploring activity patterns of the Taipei public bike sharing system," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 11, pp. 1012–1028, 2015.

[26] J. Froehlich, J. Neumann, and N. Oliver, "Measuring the pulse of the city through shared bicycle programs," in *Proceedings of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems-UrbanSense 08*, Raleigh, NC, USA, November 2008.

[27] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding bike-sharing systems using data mining: exploring activity patterns," *Procedia - Social and Behavioral Sciences*, vol. 20, pp. 514–523, 2011.

[28] N. Lathia, S. Ahmed, and L. Capra, "Measuring the impact of opening the London shared bicycle scheme to casual users," *Transportation Research Part C: Emerging Technologies*, vol. 22, no. 5, pp. 88–102, 2012.

[29] P. Borgnat, P. Abry, P. Flandrin, C. Robardet, J.-B. Rouquier, and E. Fleury, "Shared bicycles in a city: a signal processing and data analysis perspective," *Advances in Complex Systems*, vol. 14, no. 3, pp. 415–438, 2011.

[30] O. O Brien, J. Cheshire, and M. Batty, "Mining bicycle sharing data for generating insights into sustainable transport systems," *Journal of Transport Geography*, vol. 34, pp. 262–273, 2014.

[31] C. Etienne and O. Latifa, "Model-based count series clustering for bike sharing system usage mining: a case study with the vélib' system of paris," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, 2014.

[32] X. L. Zhou, "Understanding spatiotemporal patterns of biking behavior by analyzing massive bike-sharing data in Chicago," *PLoS One*, vol. 10, no. 10, 2015.

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

[34] S. Reiss and K. Bogenberger, "GPS-data analysis of munich's free-floating bike sharing system and application of an operator-based relocation strategy," in *Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems-(ITSC 2015)*, September 2015.

[35] A. Pal, Yu Zhang, and C. Kwon, *Analyzing Mobility Patterns and Imbalance of Free-Floating Bike Sharing Systems*, Transportation Research Board, Washington, DC, USA, 2017.

[36] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Addison-Wesley, Boston, MA, USA, 2006.

[37] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, 1975.

[38] D. Comaniciu and P. Meer, "Mean shift: a Robust approach towards feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, 2002.

[39] M. Fernández-Delgado and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems," *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[40] L. Breiman, L. Breiman, and R. A. Cutler, "Random forests machine learning," *Journal of Clinical Microbiology*, vol. 2, pp. 199–228, 2001.