



Application of Cluster Analysis with Unsupervised Learning to Dockless Shared Bicycle Flow Control and Dispatching

Shang-Yuan Chen¹ , and Tzu-Tien Chen² 

¹School of Architecture, Feng Chia University

²The University of Hong Kong

Corresponding author: Shang-Yuan Chen, shangyuanc@gmail.com

Abstract. After dockless shared bicycles are introduced on a large scale in the city, supply and demand problems involving different areas often emerge. This study applied cluster analysis with unsupervised learning to shared bicycle flow control and dispatching, and employed artificial intelligence to extract real, full-scale transportation rules from open data. First, this study proposes a model of the shared bicycle control system and an incentive mechanism for the reverse flow of bicycles based on threshold values. Then, we use the kernel density spatial clustering method to perform partitioning, grading, and incentives of check-out and check-in points' density in the area, furthermore, adopt the DBSCAN clustering method to establish dispersal and dispatching strategies. This study uses Shanghai Open Data in modeling, verification, and used Rstudio software to produce visualized interactive graphics for demonstration.

Keywords: K-means clustering, kernel density spatial clustering, density-based spatial clustering of applications with noise.

DOI: <https://doi.org/10.14733/cadaps.2020.1067-1083>

1 MOTIVATION AND GOAL

The promotion of shared bicycles seeks to resolve the problem of short-distance urban transportation, and shared bicycles are an important means of contemporary green transportation [3]. The docking shared bicycles found throughout Taiwan are chiefly aimed at travelers. But because, unlike dockless shared bicycles, docking shared bicycles do not allow users to obtain and return the bikes at any location, docking shared bicycles cannot easily become important means of everyday transportation for urban residents. And although dockless V-bikes [6] made a short-lived appearance in Taiwan during 2017, little data concerning V-bikes is available. As a consequence, this study relies on observations and analysis of dockless shared bicycles in Shanghai, China to gain a better understanding of dockless shared bicycles and methods of improving their effectiveness, and specific methods of making them more acceptable to city residents and a more important means of transportation.

After shared bicycles are introduced on a large scale in the city, supply and demand problems involving different areas often emerge. For instance, at certain important transportation nodes, bicycles frequently build up during peak times, and may also become scarce. In addition, when dispatching is used reactively to resolve the problems of bicycle oversupplied and scarcity, this approach is often too slow to be of much use. Furthermore, this approach also imposes logistics costs on bicycle companies, which are required to use dispatching manpower and trucks to move bicycles from areas where they are in oversupply to areas where they are scarce. In addition, because bicycles cannot be used while they are being transported, this reduces urban transportation capacity. As a consequence, the dispatching of shared bicycles in real-time, or in advance, is especially important.

This study applied cluster analysis with unsupervised learning to shared bicycle flow control and dispatching, and employed artificial intelligence to extract real, full-scale transportation rules from open data through the application of cluster analysis with unsupervised learning. This study's analysis of the times and locations at which the oversupply or shortage of shared bicycles may occur in Shanghai will help bicycle companies to perform dispatching ahead of time. Furthermore, clustering results can be used in conjunction with incentive mechanisms employing threshold values to encourage bicycle users to actively perform bicycle dispersal and dispatching, which will boost the carrying capacity of urban transportation systems, and reduce the risk and cost of ineffective dispatching.

2 LITERATURE REVIEW

The scope of this study includes shared bicycles and cluster analysis; the following is a review of the literature:

2.1 Shared Bicycles

Shared bicycle systems include docking and dockless models. The world's first intelligent public bicycle system was the Vélo'v docking public bicycle system, which was established in Lyon, France during 2005 [19]. Nevertheless, docking shared bicycles have never caught on in China. Since the people of China consider bicycles to be everyday means of transportation, the vast flows of bikes that occur in pace with the rhythms of people's daily work and activities would tend to cause major problems for docking shared bicycles. Since the number of docks at a docking station is fixed, the number of bicycles will exceed the number of docks at peak times, and when such spillover occurs, the only solution is to use limited manpower and trucks to transport and dispatch the bicycles, and such dispatching will inevitably be too late to meet real-time need. Many dockless shared bicycles began appearing in China during the second half of 2016. Compared with docking shared bicycles, dockless bicycles are much more versatile and convenient; they are not dependent on fixed, managed parking areas, and can be left in any vacant space along roads. Consumers need only use Wechat Pay, Alipay, or some other app to pay a certain deposit before using the bikes. The following is a review of the clustering principles of the two shared bicycle models and the differences between them:

2.1.1 Docking shared bicycles and their clustering principles

In order to predict the number of bicycles at each group of docks in a city as a whole and after clustering, it is first necessary to perform clustering of the docking points [11]. In the case of docking shared bicycles, because the locations at which bicycles can be obtained or returned are limited to places where docks are present, it is only necessary to perform clustering of the fixed docks, instead of the bicycles themselves. As a consequence, it is much easier and more clear-cut to calculate the number of docking bicycles that are taken or returned than dockless bicycles. If, within a certain geographical scope, we divide n docks into m groups ($n > m$), calculate the probability of bicycles moving among the m groups, and analyze their historical data (including time, climate, notable events, and other influencing factors), we can then predict future movement trends. In Figure 1,

after dividing docks into 4 groups (clusters) (G1-G4), we have modeled the probability of bicycles moving between groups. In this diagram, the arrows represent the process of directional movement from a starting point (check-out) to an ending point (check-in), and the numbers represent probabilities. The sum of the check-out probabilities indicated by the arrows beginning at any one group will be 1; for instance, the sum of group G1's four originating probabilities is 1 ($0.1+0.5+0.39+0.01=1$), and the same for groups G2-G4.

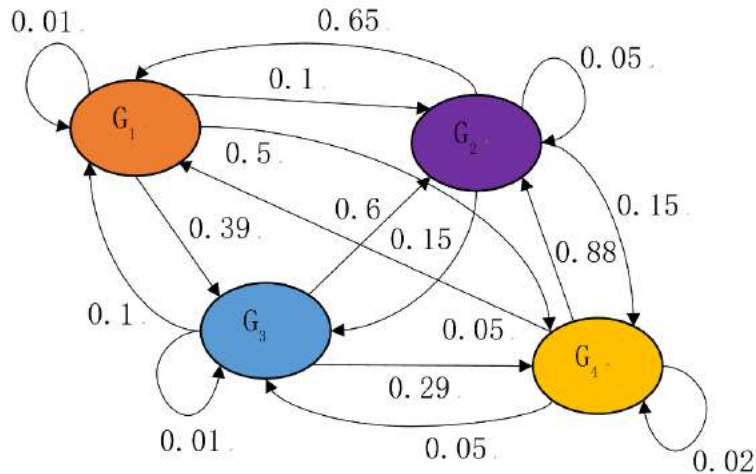


Figure 1: Probability of bicycles moving between groups (adapted from Li, 2015).

2.1.2 Dockless shared bicycles and clustering principles

Unlike the case of shared bicycles with fixed docks, the clustering of dockless shared bicycles changes with time, and it is difficult to calculate check-out and check-in probabilities. One feasible approach that can be employed to cluster dockless shared bicycles is to perform clustering of the coordinates of the bicycles' check-out and check-in locations, but in contrast to the clustering of fixed docks, this is a much more difficult task. The fact that check-out points and check-in points have an irregular, macro distribution makes it very hard to perform direct clustering. Moreover, the clustering of dockless shared bicycles must also reflect the precondition that users within each cluster must be able to easily find usable shared bicycles within a comfortable walking distance. In addition, the most suitable walking distance is also affected by climatic conditions, vertical distance, and a city's state of economic development [7]. In the case of Shanghai community residents, taking a walking distance of 787m as a dividing point, residents will choose to walk more than 10% of the time when the distance is closer than this, but will very seldom choose to walk when the distance is greater than 787m [18]. This study, therefore, adopted the round figure of 700m as the scope of the most suitable walking distance; 700m serves as the baseline scope of clusters in the remainder of this paper, and since 700m is roughly equivalent to the distance that can be walked in 10 minutes, it is representative of local everyday walking.

2.2 Cluster Analysis

As described above, the methods and algorithms used to cluster shared bicycles will affect shared bicycle flow prediction and dispatching ability. In machine learning, the use of clustering methods and the derivation of clustering algorithms are referred to as "cluster analysis." Cluster analysis is defined as the division of heterogeneous objects into homogeneous subgroups [12]. Cluster analysis can be applied to many different fields, including machine learning, data mining, model identification, image analysis, and bio-information. Clustering involves use of a static classification method to assign

similar objects to different groups or subsets, and ensures that the members of any one subset have certain similar attributes.

Data clustering is ordinarily classified as unsupervised learning. While clustering and classification are easily confused, particularly in data mining, the difference between clustering and classification lies in the fact that clustering is a process of grouping to form samples, where the samples are adjacent in space. In contrast, classification involves attaching labels to samples, where the labels are derived from external standards. In other words, clustering constitutes an unsupervised grouping process, where groups are formed naturally after data has been input [13], but the classification is supervised and involves other external input. Unsupervised neural networks are able to uncover characteristics that are important, or that may be overlooked, from data on their own, without any preconceived expectation of the output values, and can, therefore, perform clustering of data [16, 8]. Gan created the parsing tree shown below in Figure 2 based on the characteristics of clustering algorithms [5]. From the top, this tree lists hard clustering and fuzzy clustering, where hard clustering includes partitional and hierarchical approaches, and hierarchical approaches may consist of either divisive or agglomerative methods. In accordance with the nature of the dockless shared bicycle problem, this study employed hard clustering as a clustering incentive mechanism and investigated the most commonly used unsupervised learning, hard clustering, and partitioning algorithms and their applicability.

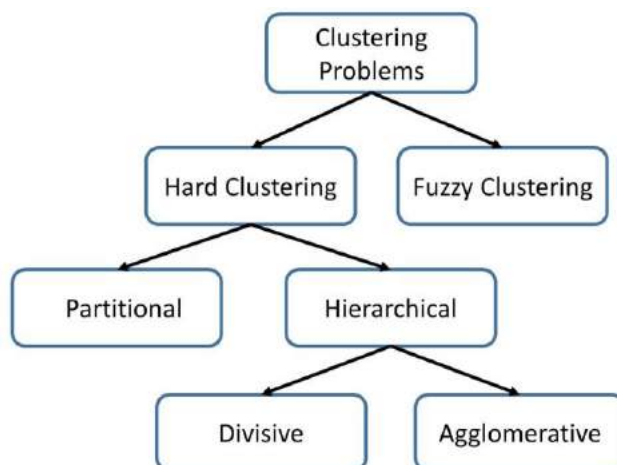


Figure 2: Parsing tree of clustering methods [5].

3 THEORY AND METHOD

As mentioned above, this study proposes a "shared bicycle control system," and applies cluster analysis with unsupervised learning to bicycle flow control and dispatching. This study also compares and analyzes three types of algorithms commonly used in hard clustering and partitioning: (1) K-means clustering, (2) kernel density spatial clustering, and (3) the density-based spatial clustering of applications with noise (DBSCAN). This study employed the R language in applications program development and used the integrated R development environment provided by RStudio software.

3.1 Shared Bicycle Control System

It is extremely important that shared bicycle companies possess real-time or ahead of time dispatching and management ability. This study proposes a shared bicycle control system model and an incentive mechanism for the reverse flow of bicycles based on threshold values. The clustering results can be used in conjunction with incentive mechanisms employing threshold values to

encourage bicycle users to actively perform bicycle dispersal and dispatching, which will boost the carrying capacity of urban transportation systems, and reduce the risk and cost of ineffective dispatching. As shown in Figure 3, this system considers the number of check-outs within a unit time and a unit area to be the bicycle demand of that area, and the number of check-ins within a unit time and a unit area to be the bicycle supply of that area. In order to ensure that each area has a suitable amount of bicycle stock (making it a "normal area" between the upper and lower points at which incentive mechanisms are triggered in the diagram below), incentives or compulsory rules must be used to control the rate of check-outs and check-ins. When the stock is excessively high—higher than the upper threshold—the system will trigger the check-out incentive mechanism. When the stock is too low—below the lower threshold—the system will activate the check-in incentive mechanism. These incentive mechanisms controlling reverse flow based on threshold values can be expected to ensure that stock will be in a better state than when no incentive mechanisms are used. Apart from this, when the system detects that the bicycle density has reached the level of oversupply or shortage, the shared bicycle company will have to dispatch trucks to perform transport and force the system back within the normal range.

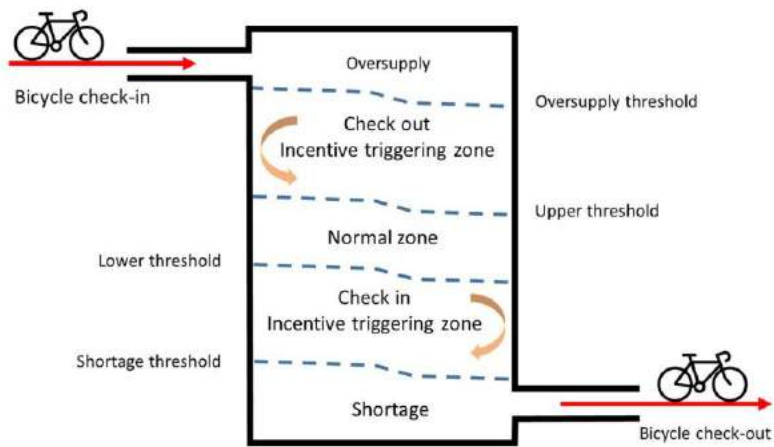


Figure 3: Shared bicycle control system model.

3.2 The Three Types of Cluster Analysis Method

The following is an analysis of three types of commonly used algorithms for unsupervised learning in hard clustering and partitioning, and assessment of the situations in which they can be used:

3.2.1 *K*-means clustering

K-means clustering consists of the following steps: n points are divided into K clusters, where K is a whole number, and the use of data mining to find clusters in which the sum of the squares of the distance between the data points in each cluster and the cluster center is a minimum, which is expressed by Formula 1 [14]:

$$f(x) = \sum_{j=1}^k \sum_{i=1}^{n_j} \|X_i^j - C_j\|^2 \quad (1)$$

k : number of clusters

n_j : number of data points in the j^{th} cluster

X_i^j : the i^{th} data point in the j^{th} cluster

This algorithm is composed of the following steps:

- I. K points are placed in a space expressed by the clustering subjects. These points represent the initial cluster centroids.

- II. Each subject is assigned to the cluster with the closest centroid.
 - III. After all, subjects have been assigned, re-calculate the locations of the K centroids.
 - IV. Repeat steps 2 and 3 until the minimum sum of the square of the error (i.e., distance) of the points in each cluster and the centroid is found, at which time the centroids will no longer move.
- The left side of Figure 4 shows clustering in a situation where $K=3$, which yields the results shown on the right side of Figure 4.

While K-means clustering can yield excellent clustering results in cases where clusters have a quasi-spherical shape and differences in the distance are not great, but cannot readily reveal clustering in cases involving arbitrary shapes and large differences in the distance; this method is also very sensitive to noisy data [1].

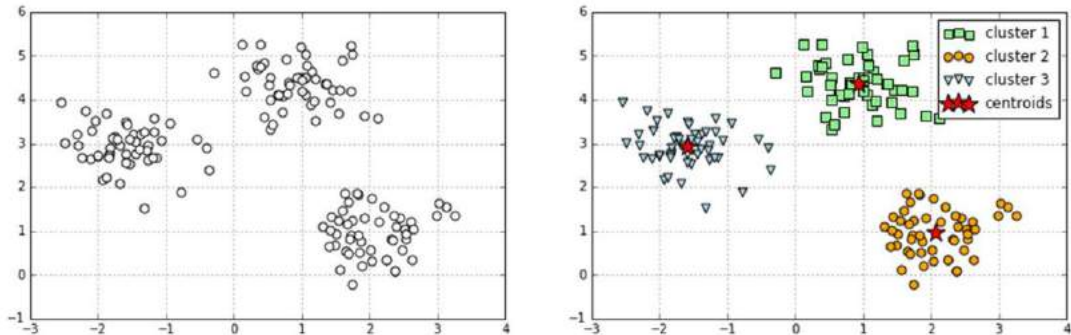


Figure 4: Example of K-means clustering.

3.2.2 Kernel density spatial clustering

Kernel density spatial Clustering is based on density levels, where areas with a high density of objects are surrounded by relatively low-density areas [5]. This section uses the density clustering method known as the "kernel density estimation" method; this method involves the use of the known locations of discrete events to estimate the overall density of a two-dimensional research domain. This estimated density is used to cover the top of the research domain with a grid cell, and the estimation of density is performed based on each grid cell's center point (Figure 5). The weighting of each distance between a discrete event and grid cell center point is then performed employing a kernel function and bandwidth [17] (Figure 6). In other words, the nature of the optimization and generalization function is a major characteristic of this density estimation method, which can perform fitting and generalization of large bodies of complex data in accordance with the approximation function and bandwidth, and possesses estimation ability. Commonly used density functions include such as rectangle, Epanechnikov curve, and Gaussian curve.

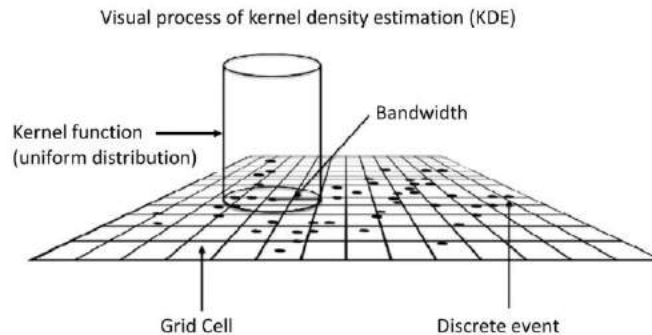


Figure 5: Performing density estimation based on the center point of each grid cell [17].

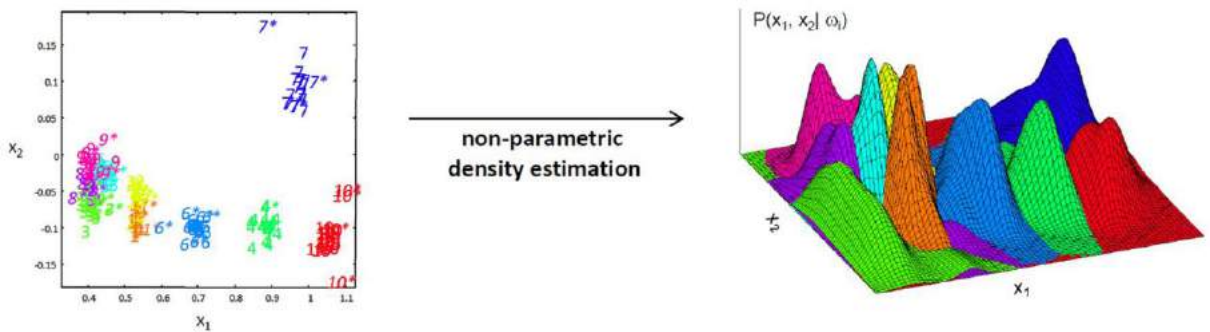


Figure 6: Weighting based on a specific kernel function and bandwidth [9].

3.2.3 Density-based spatial clustering of applications with noise (DBSCAN)

The DBSCAN method is also based on density: When a set of points is placed in a given space, this algorithm can assign nearby points to individual groups, and mark points located in low-density areas as outside points, which constitute noise points, and are not connected with any high-density areas. In Figure 7, point n is a noise point. The goal of DBSCAN is to find the largest set of density-connected objects; in Figure 8, the areas around data points p , q , and t are mutually connected.

The following are key definitions employed in the DBSCAN clustering method [4]: 1. The adjacent area within radius ϵ of data point p is known as that data point's ϵ -adjacent area. 2. A data point's ϵ -adjacent area must contain at least a minimum number of data points (MinPts), and data point p constitutes a kernel object (Figure 9). 3. When the location of data point p is within the ϵ -adjacent area of a certain kernel object q , data point p is termed directly density-reachable object of q . 4. Assuming that data point p is directly density-reachable from q_1 , q_1 is directly density-reachable from q_2 , and q_{i-1} is directly density-reachable from q_i , then data point p is termed density-reachable from q_i . 5. Assuming that data points p and q are both density-reachable from o , then p and q can be said to be density-connected. As shown in Figure 10, when control of the scanning radius and the minimum number of points minimize the noise in the ultimate clustering results, this is generally felt to be an optimal DBSCAN result.

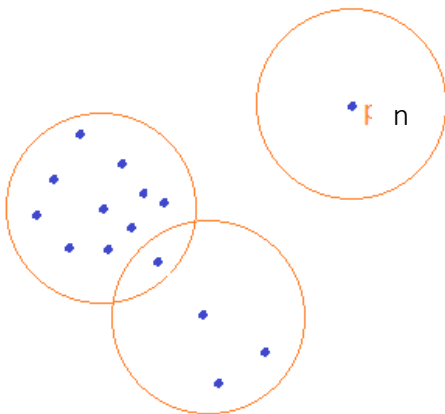


Figure 7: Point n is a noise point.

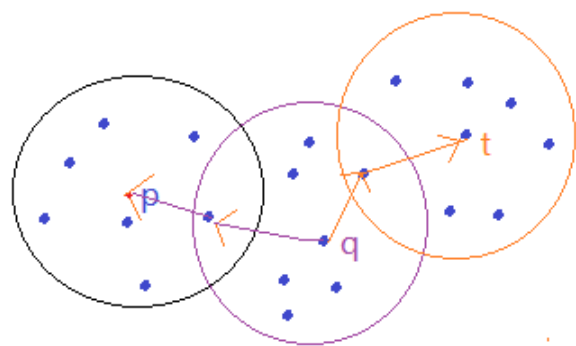


Figure 8: Finding the largest set of density-connected objects.

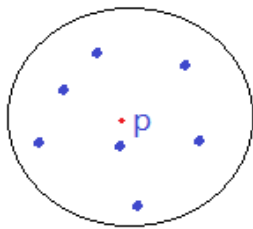


Figure 9: Point p is a kernel object.

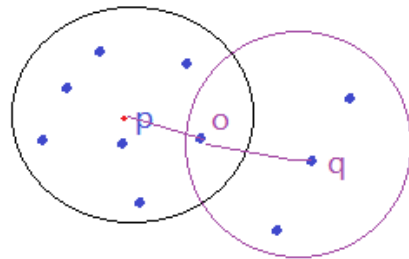


Figure 10: p and q are density-reachable.

Both types of density-based algorithms, K-means clustering, and DBSCAN, do not require the prior determination of the number of clusters, which is because these density-based clustering algorithms are able to automatically detect clusters, and naturally, divide data into a number of clusters. In addition, in comparison with other methods, these two methods yield excellent results when clusters have arbitrary shapes. As a consequence of their advantages, these density-based clustering algorithms are very suitable for analysis of the irregular distribution of dockless shared bicycles in cities.

Comparing the kernel density spatial clustering and DBSCAN methods, which are both density-based algorithms, the former offers the advantage of easy density estimation and can be used to perform partitioning and grading. However, the latter can only be used to mark those clusters that meet a preset density requirement. However, due to the need to set a MinPt parameter, DBSCAN can effectively avoid the single-link effect, which occurs when different clusters are connected by a single point or relatively few lines, which causes them to be seen as a single cluster. Because of this, the control radius ϵ and MinPts quantity can be used to create a control threshold. As a consequence, this study will use these two types of density-based clustering methods in subsequent applied research.

3.3 Basic Tools and Program Language

This study used the R programming language as a data processing and analysis tool. The R language chiefly employs command line commands, uses call functions to perform data processing and analysis, and displays output in a graphic program. Rstudio is an integrated development environment for R language programming and is an open-source program. Rstudio has many coding tools and project management interfaces and enables the convenient maintenance and rapid browsing of files and function sets. In addition, it can also package code written in the R language, enabling it to produce visualized interactive graphics [10]. (Figure 11)

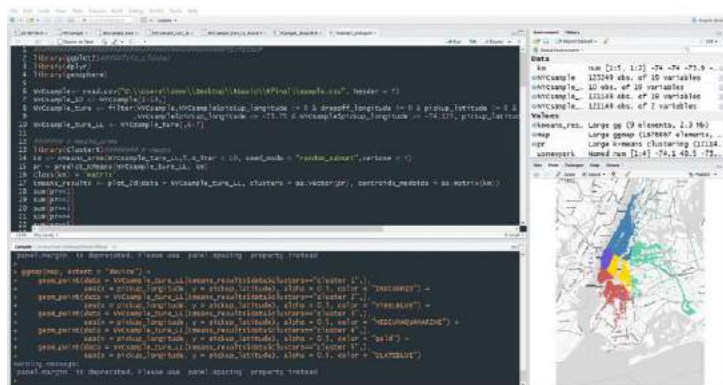


Figure 11: Rstudio's user interface.

4 REALIZATION AND VERIFICATION

This section realizes and tests a shared bicycle control system in accordance with the foregoing theory and method. We first perform scenario modeling of the incentive mechanisms controlling reverse movement based on threshold values in order to gain an understanding of problems that might be faced in actual operation. Afterwards we employ 101,843 data items issued in August 2016 for the Shanghai Open Data Innovative Application Competition [15] to perform modeling and verification.

4.1 Scenario Modeling Incentive Mechanisms Controlling Reverse Movement Based on Threshold Values

Within any one cluster, the distance users needed to go to find shared bicycles did not exceed the most suitable walking distance, which implied that the minimum length of the diagonal lines across grid squares must be 700m, and the sides of the squares were 550m. (Figure 12) This section expresses the unit density of bicycles within clusters in terms of numeric values. We assumed a state of balance and normal stock at the initial time t_0 (00:00 on 8/1/2016). The bicycle density within a unit area was 6 units in all cases (left side, Figure 13). After one hour, (01:00, designated t_1) we calculated changes in the unit density of the bicycles stock (right side, Figure 13). As a consequence, we were able to calculate the bicycle check-out rate (left side, Figure 14) and check-in rate (right side, Figure 14) of each grid square at a time t_0 . (The areas with underlined numbers in these figures have a check-out rate \geq check-in rate, which implies that $X_{t_{n-1}} \geq Y_{t_{n-1}}$ when $n=1$)

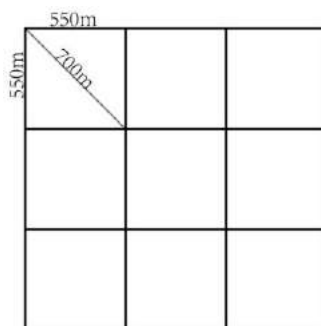


Figure 12: Schematic diagram of scenario modeling of clustering.

The situations at t_0 and t_1 on 8/1/2016 were:

6	6	6	6	<u>5</u>	6
6	6	6	<u>5</u>	10	<u>5</u>
6	6	6	6	<u>5</u>	6

Figure 13: Changes in unit bicycle density in clusters: Initial state (left) and state at time t_1 (right).

Check-out			Check-in		
-1	<u>-2</u>	-2	+1	<u>+1</u>	+2
<u>-3</u>	-2	<u>-2</u>	<u>+2</u>	+6	<u>+1</u>
-2	<u>-2</u>	-4	+2	<u>+1</u>	+4

Figure 14: Check-out rate (left) and check-in rate (right) of bicycles in each cluster at time t_0 .

In accordance with the foregoing observations, the most important aspect is maintaining a balance between supply and demand within each cluster in this dynamic system, and incentives must seek to reverse imbalances in supply and demand during the previous period of time. This study consequently designed check-out incentive formula (2) and check-in incentive formula (3) (in which X is the check-out rate, Y is the check-in rate, t_n is the time, and n is an integer):

$$\text{I. If } X_{t_{n-1}} \leq Y_{t_{n-1}} \text{ at time } t_{n-1}, \text{ then } \frac{|Y_{t_{n-1}} - X_{t_{n-1}}|}{2^i} \text{ is the check-out incentive amount (2)}$$

Where i is a whole number and is the check-out sequence at time t_n .

In accordance with formula (2), at time t_n , the incentive for check-out of the first bicycle is the largest, and the incentive for the check-out of each subsequent bicycle will decrease progressively with the increasing root of $\frac{1}{2}$ with the increase in the number of bicycles checked out: $\frac{1}{2^1}, \frac{1}{2^2}, \frac{1}{2^3}, \dots$ and so on. The incentive for the final bicycle checked-out will approach 0 without limit, but will never reach 0.

$$\text{II. If } X_{t_{n-1}} > Y_{t_{n-1}} \text{ at time } t_{n-1}, \text{ then } \frac{|X_{t_{n-1}} - Y_{t_{n-1}}|}{2^j} \text{ is the check-in incentive amount (3)}$$

Where j is a whole number and is the check-in sequence at time t_n , the incentive for check-in of the first bicycle is largest, and the incentive for the check-in of each subsequent bicycle will decrease progressively with the increasing root of $\frac{1}{2}$: $\frac{1}{2^1}, \frac{1}{2^2}, \frac{1}{2^3}, \dots$ and so on. The incentive for the final bicycle checked-in will approach 0 without limit, but will never reach 0.

As described above, if a bicycle is checked out from area p at time t_{out} and checked in area q at t_{in} , the incentive amount will be the sum of the check-in incentive in area p during at time t_{out} and the check-out incentive in area q during at time t_{in} . This can be represented as:

$$\text{Sum} = \frac{|Y_{t_{out(n-1)}} - X_{t_{out(n-1)}}|}{2^i} + \frac{|X_{t_{in(n-1)}} - Y_{t_{in(n-1)}}|}{2^j}, \text{ which is the total incentive amount (4)}$$

Based on scenario modeling of the use of threshold values to promote reverse movement, this section investigates incentive mechanism strategies and their reasonableness and establishes a prototype incentive mechanism and operating principles. Our analysis proceeded as follows: (1) although the shared bicycle control system has a normal stock range, this range is difficult to define, which is because dockless shared bicycles lack fixed docks that can be used to estimate reasonable quantities of stock and their probability of movement. Compared with docking bicycles, determining static stock ranges for dockless shared bicycles is very difficult. (2) Addressing differences in supply and demand in different clusters, dockless bicycle systems can see the number of check-outs in a unit area during a unit time as people's demand for bicycles in that area, and the number of check-in in a unit area during a unit time as the supply of bicycles in that area. Within the permissible stock scope, when there is an imbalance between supply and demand, incentive mechanisms promoting reverse movement should be implemented to ensure that the supply and demand situation more closely approaches a state of balance. (3) In order to encourage interest in user dispatching, incentives that decrease gradually with the number of bicycles should be established. The incentive amount should be directly proportional to the bicycles stock within an area and the deviation of the number of bicycles in that area from a state of balance. (4) If an area is in a peak usage period, the flow rate is fast (i.e., the pick-up rate is fast, and the return speed is also fast). Alternatively, if the relative time is in the off-peak period, the flow rate is slow (that is, the vehicle pick-up rate is slow, and the vehicle speed is also slow). When the absolute value of the difference between the check-out rate and the check-in rate is in a range, the number of bicycles in the area is considered to be in a state in which normal operation can be maintained. However, if an area is in a fast check-out speed for a long time, and the check-in speed is very slow, or almost stationary, it means that the number of bicycles in the area has been lower than the shortage threshold. Conversely, if the check-in rate is rapid for a prolonged period within a certain area, but the check-out rate is prolonged, or almost 0, this indicates that the number of bicycles in that area will quickly rise above the oversupply

threshold. When states of oversupply or shortage occur, trucks should be dispatched to move and redistribute the bicycles, which will ensure that the system returns to a normal state.

4.2 Use of Shanghai Open Data in Modeling and Verification

This study used 101,843 data items issued in August 2016 for the Shanghai Open Data Innovative Application Competition (Soda Open data, 2018) to perform modeling; these data items included order numbers, bicycle serial numbers, user serial numbers, check-out times, check-out longitudes, check-out latitudes, check-out times, check-in longitudes, check-in latitudes, and routes. We used this data to find optimal clustering methods for shared bicycles in Shanghai by examining relatively small areas (the area around Shanghai's Jingan Temple) and relatively large areas (the part of Shanghai within the fourth ring roads—121.75°E-121.125°E, 30.95°N-31.45°N), and also established models corresponding to the state of clustering. We employed the following four steps in this process: (1) Compilation and processing of the needed data, (2) data visualization analysis, (3) clustering, grading, and establishment of incentives, and (4) establishment of dispersal and dispatching strategies.

4.2.1 Stage 1: Compilation and processing of data

As of August 2016, it was projected that 223,000 bicycles of Mobike, a well-known company providing a bike sharing service, were in use in Shanghai [15]. Since each Mobike bicycle in China was ridden by 5.4 persons per day [2], we can roughly estimate that there should have been approximately 36,000,000 use records each month for Mobike bicycles in Shanghai during August 2016. Based on the Shanghai area's latitude and longitude, and the fact that 111,000m separates lines of latitude and longitude, as well as a maximum suitable walking distance of 700m, which implies grid squares with sides 550m in length, we can estimate that there was an average of approximately 2,857 bicycles in each 550m x 550m grid square at that latitude and longitude (discarded, damaged, and under-repair bicycles were not subtracted). Assuming that 50% of bicycles in the urban area of Shanghai could still be found, we can estimate approximately 1,428 bicycles in each grid square. The following modeling, therefore, employed an average stock of 1,428 bicycles in each 550m x 550m grid square as the reference bicycle stock value for Shanghai during August 2016.

4.2.2 Stage 2: Data visualization and analysis

Data visualization and analysis were performing using the RStudio software, as shown below:

The following are bicycle use and return records and analysis on sunny days suitable for bicycle use:

- I. Taking 6 hours as the basic period, people's everyday activities ensured that bicycle density underwent various changes. Taking check-outs as an example, as shown below, on Monday, the rest period from 00:00 to 06:00 (Figure 15) had fewer check-outs than the three periods from 06:00 to 12:00 (Figure 16), from 12:00 to 18:00 (Figure 17), and from 18:00 to 24:00 (Figure 18), and the number of check-outs fluctuated within these three subsequent periods.
- II. Taking 2 days as a period, and number of check-outs as an example the scope of bicycle use changed slightly between working days and rest days. As shown in figures 19-20, the scope activity was greater, and spread slightly further out into suburban areas, during rest days (Saturday and Sunday) than during working days (Thursday and Friday).
- III. Differences in the intensity of bicycle use and return occurred between different areas. As shown in figures 21-22, differences in the check-out density existed in different areas on both rest days (Saturday and Sunday) and working days (Thursday and Friday).

4.2.3 Stage 3: Application of kernel density spatial clustering, partitioning, grading, and incentives

As described above, the activity of the dynamic bicycle control system was not evenly distributed but varied in intensity and scope in accordance with people's everyday activities. Because the system's most important function was to maintain a balance between supply and demand in clusters with different density, incentives sought to reverse supply and demand imbalances during earlier

periods of time. This brings up the question of how to perform partitioning and grading of bicycle check-out points and check-in points.

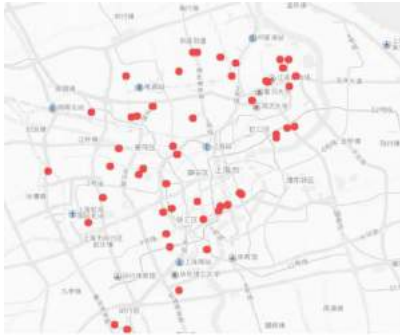


Figure 15: Schematic diagram of check-out points in Shanghai on August 1, 2016 (Monday) from 00:00 to 06:00.



Figure 16: Schematic diagram of check-out points in Shanghai on August 1, 2016 (Monday) from 06:00 to 12:00.

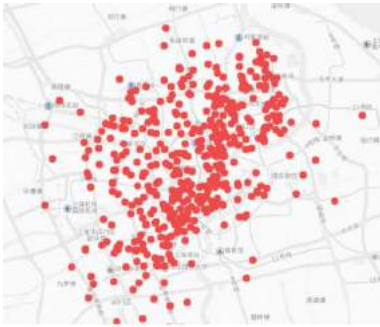


Figure 17: Schematic diagram of check-out points in Shanghai on August 1, 2016 (Monday) from 12:00 to 18:00.



Figure 18: Schematic diagram of check-out points in Shanghai on August 1, 2016 (Monday) from 18:00 to 24:00.



Figure 19: Schematic diagram of check-out points in Shanghai on August 4 & 5, 2016 (Thursday and Friday).



Figure 20: Schematic diagram of check-out points in Shanghai on August 6 & 7, 2016 (Saturday and Sunday).



Figure 21: Schematic diagram of check-out point use intensity in Shanghai on August 4 & 5, 2016 (Thursday and Friday).



Figure 22: Schematic diagram of check-out point use intensity in Shanghai on August 6 & 7, 2016 (Saturday and Sunday).

When check-outs and check-ins spanned different clusters, how should incentives be provided in order to reverse bicycle movement trends? The following is an explanation of how this was done:

I. Partitioning and grading:

In accordance with the analysis in section 3.2, we used the kernel density spatial clustering method to perform partitioning and grading of check-out and check-in density in the area around Shanghai's Jingan Temple on 8/1/2016.



Figure 23: Partitioning of check-out density in the area around Shanghai's Jingan Temple on 8/1/2016.

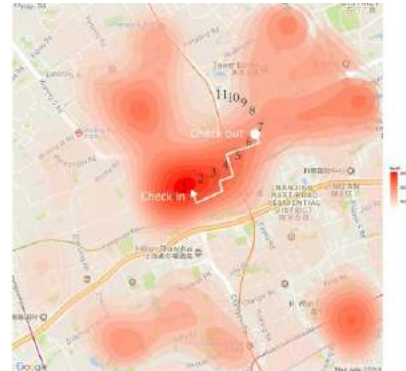


Figure 24: Partitioning of check-in density in the area around Shanghai's Jingan Temple on 8/1/2016.

■ Partitioning and grading of check-out density:

As shown in Table 1, according to the results of calculations, the level 1 area around Jingan Temple had a check-out density of approximately 3,600, while the adjacent peripheral area had a check-in density of approximately 900. This shows that the density difference between the central area and the adjacent peripheral area was approximately $3,600 - 900 = 2,700$. In accordance with the depth of the purple color in Figure 23, the system recommended that there be 12 clusters; when 2,700 points were divided among levels 1 to 12, we obtained the clustering thresholds for each level shown in Table 1:

Areas with different check-out clustering levels	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	Level 10	Level 11	Level 12
Upper threshold	≥3600	3375	3150	2925	2700	2475	2250	2025	1800	1575	1350	1125
Lower threshold	3375	3150	2925	2700	2475	2250	2025	1800	1575	1350	1125	≤900

Table 1: Check-out clustering thresholds for each level.

- Check-in density partitioning and grading:
 As shown in Table 2, according to the results of calculations, the level 1 area around Jingan Temple had a check-in density of approximately 3,100, while the adjacent peripheral area had a check-in density of approximately 800. This shows that the density difference between the central area and the adjacent peripheral area was approximately 3,100-800=2,300. In accordance with the depth of the purple color in Figure 24, the system recommended that there be 11 clusters; when 2,300 points were divided among levels 1 to 11, we obtained the clustering thresholds for each level shown in Table 2:

Areas with different check-in clustering levels	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	Level 10	Level 11
Upper threshold	≥3100	2891	2681	2472	2263	2055	1845	1636	1427	1218	1009
Lower threshold	2891	2681	2472	2263	2055	1845	1636	1427	1218	1009	≤800

Table 2: Check-in clustering thresholds for each level.

II. Incentives

As shown in figures 23-24, during the period from 12:00 to 18:00 on 8/1/2016 (i.e., the period following 06:00 to 12:00 on 8/1/2016), the first bicycle was checked-out from the check-out point and returned at the check-in point half an hour later, which was during the same time period.

- With regard to check-outs, the estimated check-out rate X at the check-out point was 1600, and this area was in a level 7 check-in cluster (threshold: 1,845-1,636); since $1600 < 1636$ (implying that the check-out rate X was lower than the minimum estimated check-in rate), consequently when $X_{min} \leq X_{t_{n-1}} \leq X_{max}$, $X_{t_{n-1}} < Y_{min}$, the check-out incentive system should be activated. According to formula (2), $Y_{t_{n-1}}$ is entered as Y_{min} ,

$$\frac{|Y_{t_{n-1}} - X_{t_{n-1}}|}{2^i} = \frac{|1636 - 1600|}{2^1} = 18$$

- With regard to check-ins, the estimated check-in density Y of the check-in point was 2,800; and this area was in a level 2 check-out cluster (threshold: 3,375~3,150); since $2750 < 3150$ (implying that the check-in rate Y was lower than the minimum estimated check-out rate), when $Y_{min} \leq Y_{t_{n-1}} \leq Y_{max}$, $X_{min} \geq Y_{t_{n-1}}$ check-in incentive system should be activated. According to formula (3), when $X_{t_{n-1}}$ is entered as X_{min} ,

$$\frac{|X_{t_{n-1}} - Y_{t_{n-1}}|}{2^j} = \frac{|3150 - 2750|}{2^1} = 200$$

- The total incentive is the sum of the two amounts. According to formula (4), a total incentive amount of 218 units is obtained.

$$\frac{|Y_{t_{out(n-1)}} - X_{t_{out(n-1)}}|}{2^i} + \frac{|X_{t_{in(n-1)}} - Y_{t_{in(n-1)}}|}{2^j} = 218$$

The foregoing is a discussion of partitioning and grading and the use of an incentive mechanism relying on threshold values to reverse the flow of bicycles. The data for the period from 12:00 to 18:00 on 8/1/2016 obtained from on-site observations was found to be very consistent with the results of our model. Because Monday is the first working day of the week, it is easy to find shared bicycles in the area around Jingan Temple, but since the Jingan Temple area contains many offices and businesses, and has intensive check-out and check-in activity, is typically not easy to find shared bicycles in this area.

4.2.4 Stage 4: Establishment of a dispersal and dispatching strategy

The DBSCAN clustering method can be used in conjunction with control radius ϵ and MinPts values to find ϵ -adjacent areas and the largest sets of density-connected objects while avoiding noise. Because of this, if we could infer the bicycle control system's oversupply and shortage thresholds, we could use the oversupply and shortage thresholds as the MinPts values in the DBSCAN method, which would allow us to find clusters with an oversupply of bicycles from ϵ -adjacent areas and density-connected objects, and we could also find clusters with bicycle shortages from noise points. This would allow the dispatching of trucks to move bicycles from clusters with an oversupply to clusters with a shortage.

As in section 4.2.1, we again employed an average stock of 1,428 bicycles in each 550m x 550m grid square as the reference bicycle stock value for Shanghai during August 2016. This allowed us to infer an average stock of approximately 7,267 bicycles within a radius of 700m, which represented the maximum suitable walking distance. If the area centered on Jingan Temple is taken as the area with the greatest demand for bicycles, the upper check-out rate threshold value for each 550m x 550m grid square would be 3,600 bikes. If twice the normal stock (7,200 bicycles) is taken to indicate oversupply, then 36,640 bicycles within a 700m radius could be taken as the oversupply threshold. Using the DBSCAN clustering method, when the radius ϵ was entered as 700m and the MinPts value was set as 36,640 bicycles, we could find clusters with an oversupply of bicycles. In Figure 25, the light green color indicates all detectable bicycles within the urban area, while the overlaid blue, dark green, and orange indicate clusters with an oversupply of bicycles (more than 36,640 bicycles). By the same principle, if a shortfall of 800 bicycles in each 550m*550m grid square indicates a shortage, 4,071 bicycles in an area with a radius of 700m can be considered the shortage threshold. Employing the DBSCAN method, the radius ϵ was entered as 700m and the MinPts value was set as 4,071 bicycles. As shown in Figure 26, after subtracting the light green noise points in the brown cluster, this cluster was seen to have a bicycle shortage.

Note: Detectable bicycles within the urban area included both parked bicycles and bicycles that happened to be in use. But because the light green base map was produced during the period from 00:00 to 06:00, when few bicycles were in use, we can ignore the effect of bicycles in use at that time in this snapshot.

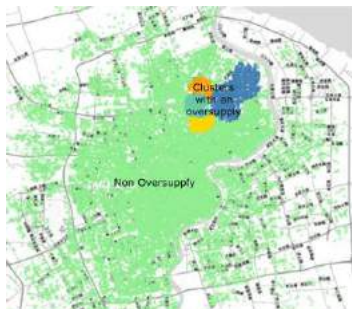


Figure 25: Analysis of areas with a bicycle density oversupply.



Figure 26: Analysis of areas with bicycle shortages.

5 CONCLUSIONS

This study established a prototype control system for dockless shared bicycles, with an incentive mechanism promoting the reverse flow of bicycles, encouraging bicycle users to autonomously maintain a balance between bicycle supply and demand. In addition, this study also applied the kernel density spatial cluster method to partitioning and grading, and developed an incentive system; and also applied the DBSCAN clustering method to finding clusters with an oversupply of bicycles and noise indicating bicycle shortages, which allowed the development of a concrete dispersal and dispatching strategy. We would also like to make the following recommendations concerning subsequent research:

- I. The accuracy of the data concerning Mobike bicycles in Shanghai was not high. In the future, with increasing data transparency and greater freedom to use data for academic purposes, accuracy can be increased further. If large amounts of data are used in training, this can enhance the clustering accuracy of research results and allowed the establishment of incentive programs with even shorter cycles, which will ensure those incentive mechanisms and dispatching strategies are even more effective.
- II. The sorting, or ordering, the capability of KDE-based Clustering helps to compare the spatial structure to determine or assign the importance of mitigation measures. It can be used for detecting traffic accident hot spots or the partitioning and grading of certain human social activities. And by setting the MinPt and radius ϵ of DBSCAN, it can be used to create a control threshold. Therefore, the relationships can be used for prediction and trend detection between spatial objects, especially a huge data set collected from different sources like satellite images, X-rays, medical images, traffic cameras, and GIS system.

S.-Y. Chen, <http://orcid.org/0000-0003-0249-5900>

REFERENCES

- [1] Aggarwal, C. C.; Reddy, C.: K. Data Clustering: Algorithms and Applications, Chapman and Hall/CRC, 2013, 88-90, 111-115, 284.
- [2] Beijing Planning Design Research Institute, 2017 Sharing Bike and Urban Development White Paper, 2017, 3-5, <http://www.199it.com/archives/581592.html>.
- [3] Chang, S. K.; Chang, H. W.; Chen, Y. W.: Green Transportation, Slow, Friendly, and Sustainable: People-Oriented Transportation Environment Makes The City Smoother and Life Better, Neoneaturalism, Taipei, 2013, 80-81. <https://www.books.com.tw/products/0010614064>.
- [4] Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, International Conference on Knowledge Discovery and Data Mining, 1996, 226-231.
- [5] Gan, G.; Ma, C.; Wu, J.: Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Mathematics, 2007, 10, <https://doi.org/10.1137/1.9780898718348>.
- [6] Inmediahk.net: Has Taiwan Shared Bicycle Success?, 2018, <https://www.inmediahk.net/node/1058497a>.
- [7] Jeff Speck, Walkable City, North Point, NY, 2012, 3-6.
- [8] Kulkarni, Arun D.: Computer Vision and Fuzzy-Neural System, Prentice Hall PTR, 2001, 351-363, 465.
- [9] L7: Kernel Density Estimation, 2018, http://research.cs.tamu.edu/prism/lectures/pr/pr_17.pdf.
- [10] Lander, J.P.: R for Everyone: Advanced Analytics and Graphics, Addison-Wesley Professional; 1 edition, 2013.

- [11] Li, Y.X.; Zheng Y.; Zhang, H.C.; Chen, L.: Traffic prediction in a bike-sharing system, 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems Seattle, 2015, No.33, 78-88.
- [12] Negnevitsky, M.: Artificial Intelligence: A Guide to Intelligent Systems, Pearson Education Limited, 2011, 303.
- [13] Príncipe, J. C.; Euliano, N.R.; Lefebvre, W.C.: Neural and Adaptive Systems: Fundamentals through Simulations, Wiley, 2000, 341-344.
- [14] Seddawy., A. B. E.; Sultan, T., Khedr, A.: Enhanced K-mean algorithm to improve decision support system under uncertain situations, International Journal of Modern Engineering Research (IJMER), 2(6), 2012, 4094-4101.
- [15] Soda Open Data, 2018, <http://shanghai.sodachallenges.com/data.html>.
- [16] Su, M. C.; Zhang, X. D.: Machine Learning: Neural Networks, Fuzzy Systems, and Gene Algorithms (second edition), Chuan Hwa, 2004, 4.2. <https://www.books.com.tw/products/0010029251>.
- [17] Timothy, H.; Zandbergen, P.: Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting, Policing: An International Journal of Police Strategies & Management, 37(2), 305-323, 2014, <https://doi.org/10.1108/PIJPSM-04-2013-0039>.
- [18] Wang, N.; Du, Y. C.: Resident walking distance threshold of community, Transportation Standardization, 2015, 2, <http://doi.cnki.net/Resolution/Handler?doi=10.16503/j.cnki.2095-9931.2015.02.004>.
- [19] Zhu, W.; He J. Y.; Wang D.: Methods and empirical research on the distribution of public bicycle systems in France: case study on Paris and Lyon, UPI, 30(SI), 2015, <http://www.upi-planning.org/Files/hjcsqh/MagazinePDF/bf7e31c9-d690-4dd0-90d5-340f0a3cc907.pdf>.