A novel partition model of scheduling regions for public bicycle system

Fei Lin¹, Yang Yang¹,³, Bin Zheng², Hong Ma¹ and ShiHua Wang¹

¹ School of Computer Science and Technology, Hangzhou Dianzi University, 310018 Hangzhou, P.R.China
² Zhejiang Economic Information Center, 310006 Hangzhou, P.R.China
³ E-mail: yangyang_hdu@163.com

Abstract. Scheduling optimization, especially public bicycle scheduling analysis, has become a hot topic due to its benefits in social, economic and environmental aspects. The traditional methods mainly focus on the distance among bicycle stations to realize a self-balancing bicycle supply by minimizing the scheduling workload for each region but ignore the origin/destination flow-data and topological features of bicycle station networks. This paper proposes a novel partition model for public bicycle scheduling region. Besides the distance, the proposed model combines dynamic origin/destination flows of bicycle stations with the topological features of them in bicycle networks to construct a similarity matrix. Then scheduling region partition is obtained by using the affinity propagation clustering algorithm. To evaluate the performance of the proposed model, we compare it with the other baseline methods by calculating the correlation coefficient between the regions while the coupling coefficient functions on the real data from Hangzhou public bicycle system. The experimental result shows that the proposed model can achieve the smallest scheduling workload among the scheduling regions.

1. Introduction

Public Bicycle System (PBS) is widely applied in an attempt to deal with the problems associated with extensive usage of private transportation such as decreasing carbon emissions, avoiding traffic congestion and wisely using non-renewable resources. The application of PBS can be traced back to the late 20th century in Europe and the U.S. [1]. Due to the random and uncoordinated movements of users, starvation (fewer than requirement) or congestion (more than requirement) of bicycles at stations happen frequently, which results in a significant loss of user satisfaction [2]. However, the traditional PBS models [3]-[5] mainly focus on scheduling and maintain the balance of bicycle allocation based on distances among stations. In fact, scheduling region partition would have a positive influence on timely scheduling and efficiently scheduling route. Therefore, the similarity of stations comes from both node attributes and topological features of nodes. Then we use the affinity propagation (AP) clustering algorithm to do partition on scheduling region. The objective function of the coupling coefficient between the scheduling regions is calculated and results are compared with other baseline methods on the real data from Hangzhou public bicycle system.

2. Related work

Given the essentiality of bicycle scheduling system, it has been studied extensively in the literature. At the beginning, study topic focuses on static scheduling. The static bicycle scheduling problem is
introduced by Benchimol et al.[6]. Chemla et al. suppose that the bicycle stock is known, and then propose an effective heuristic solution method, which aims at minimizing the routing cost [7]. Raviv et al. put forward a solution base on scalable exact and approximate algorithms to solve this problem by using variable neighbourhood search heuristic or by employing constraints from stock management literature [8]. Di Gaspero et al. solve the problem base on large neighbourhood search by employing constraint programming [9]. These solutions will be useful if the requirement pattern is stable and predictable. However, if the requirement changes over time, static scheduling is not sufficient to keep the bicycle system balance status. So people change to study the dynamic scheduling of bicycles. Shu et al. proposes an optimized model for dynamic scheduling to reduce the unmet demand [10]. All the papers in this topic suppose a known distribution of demand and they are robust to the unstable requirement scenarios. In addition to the dynamic scheduling strategy, people also try to use the prediction and analysis of requirement in PBS for scheduling decision. Nair et al. assess the service level of PBS using dual-bounded chance constraint [11]. Borgnat et al. propose the idea of predicting temporal requirement and provide users with predictive information [12], [13]. Kabra et al. use position distribution to represent the customer arrival process at base stations [14]. With respect to the facility location problem, Meilin Wen develop several optimization models based on fuzzy random demand, including the expected cost minimization model and the chance maximization model [15].

3. Scheduling region division model

3.1. Problem definition
The purpose of scheduling urban public bicycles is to maintain balance status between the resources allocated to each station [16]. There are two kinds of relationships between the bicycle stations [17]. The first is the lease/return relationship between any two stations [18], as shown in figure 1. Lease from the station $i$ and return to the station $j$ or lease from the station $j$ and return to the station $i$ indicates the existence of a rent relationship between the two stations [19]. The second relationship is the space distance between two stations, as shown in figure 2.

![Figure 1. Lease/return relationship between two stations.](image1)

In our study, we use these two relationships to establish an effective scheduling region partition model, divide the isolated public bicycle stations shown in figure 3. Base on topology network structure shown in figure 4. Into the scheduling regions shown in figure 5.

![Figure 3. Space distance relationship.](image3)

![Figure 4. Topology of stations.](image4)

![Figure 5. Region partition of bicycle stations.](image5)
3.2. A mathematical model construction

There are $n$ bicycle rental stations to be partitioned into $m$ scheduling regions. Therefore, $S = \{S_1, S_2, S_3, \ldots, S_n\}$ represents the set of all bicycle stations, and $M = \{M_1, M_2, M_3, \ldots, M_m\}$ is the set of the scheduling areas to be partitioned. Element $d_{ij}$ denotes the Euclidean distance between station $i$ and station $j$. $r_{ij}$ denotes the sum of lease/return number between station $i$ and station $j$, namely the sum of the number of returning to station $j$ leasing from station $i$ and the number of returning to station $i$ leasing from station $j$ ($i,j \in N$). Accordingly, we establish the space distance matrix $\theta$ and the lease/return matrix $y$ for all stations. Also stated above, in the traditional clustering methods, due to only considering the distance, it is not reasonable to classify the regions with special geographical features such as crossing river, crossing mountain. However, firstly all the lease/return data $r_{ij}$ must be normalized before integrated as a weighting factor $\omega_{ij}$ into the distance matrix $\theta$. This normalization process is expressed as follow:

$$\theta = \begin{bmatrix}
  d_{11} & d_{12} & d_{13} & \cdots & d_{1m} \\
  d_{21} & d_{22} & d_{23} & \cdots & d_{2m} \\
  d_{31} & d_{32} & d_{33} & \cdots & d_{3m} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  d_{n1} & d_{n2} & d_{n3} & \cdots & d_{nm}
\end{bmatrix} \quad (1)$$

$$y = \begin{bmatrix}
  r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\
  r_{21} & r_{22} & r_{23} & \cdots & r_{2n} \\
  r_{31} & r_{32} & r_{33} & \cdots & r_{3n} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  r_{n1} & r_{n2} & r_{n3} & \cdots & r_{nn}
\end{bmatrix} \quad (2)$$

$$\omega_{ij} = 1 - \frac{r_{ij} - r_{\text{min}}}{r_{\text{max}} - r_{\text{min}}} \quad (3)$$

$$S_n = \begin{bmatrix}
  -\omega_{11} * d_{11} & -\omega_{12} * d_{12} & \cdots & -\omega_{1n} * d_{1n} \\
  -\omega_{21} * d_{21} & -\omega_{22} * d_{22} & \cdots & -\omega_{2n} * d_{2n} \\
  -\omega_{31} * d_{31} & -\omega_{32} * d_{32} & \cdots & -\omega_{3n} * d_{3n} \\
  \vdots & \vdots & \vdots & \ddots \\
  -\omega_{n1} * d_{n1} & -\omega_{n2} * d_{n2} & \cdots & -\omega_{nn} * d_{nn}
\end{bmatrix} \quad (4)$$

Where $r_{\text{max}}$ denotes the maximum value in the matrix, and $r_{\text{min}}$ denotes the minimum value in the matrix. $\omega_{ij}$ is the smaller the value $\omega_{ij}$ is. A smaller $\omega_{ij}$ indicates that the rent between the two stations is more frequently. Generate a similarity matrix $S_n$ based on $\omega_{ij}$ and $d_{ij}$: The network of bicycle stations is a topology network, and define $G = (V, E)$ be an undirected graph with vertices set $V$ and edges set $E$.

3.2.1. Degree centrality is defined as the number of links on a vertex. The degree centrality of a vertex, for a given graph $G = (V, E)$, is defined as the follow formula. The degree centrality of vertices and edges can be calculated by using dense adjacency matrix or sparse matrix representation of graph.

$$C_D(v) = \text{deg}(v) \quad (5)$$

3.2.2. Closeness centrality is an assessment of vertex proximity in a graph. It is the inverse of the sum of the shortest route distances between each vertex and every other vertices in the graph. The more central a vertex is, the closer it is to other vertices. Bavelas [20] define the closeness centrality as the reciprocal of the distance that is

$$C(x) = \frac{1}{\sum_y d(y,x)} \quad (6)$$

The $d(y, x)$ is the distance between vertices $x$ and $y$. In a real application, people usually refer to its normalized form instead of their sum, and the normalized form represents the average length of the shortest paths. It is generally divided by $N - 1$ from previous formula, where $N$ is the number of vertices in the graph. If the graph is very large, the results dividing by $N - 1$ or $N$ is approximate. So $N - 1$ will be replaced by $N$ resulting in:

$$C(x) = \frac{N}{\sum_y d(y,x)} \quad (7)$$
3.2.3. **Betweenness centrality** is a centrality measure of a vertex. Betweenness centrality represents the number of times a vertex on the shortest route between two other vertices. The betweenness of a vertex \( v \) in a graph \( G = (V, E) \) can be calculated as follows: First, for each pair of vertices \( (s, t) \), calculate the shortest routes between them. Second, for each pair of vertices \( (s, t) \), calculate the fraction of shortest routes which pass through the vertex. Third, sum value over all pairs of vertices \( (s, t) \). The betweenness can be represented more concisely as:

\[
C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}
\]  

(8)

Where \( \sigma_{st} \) represent the total number of shortest routes from vertex \( s \) to vertex \( t \) and \( \sigma_{st}(v) \) represent the total number of routes which pass through vertex \( v \). We can divide through the number of pairs of vertices not including vertex \( v \) which for undirected graphs is \( (n-1)(n-2)/2 \) to normalize the betweenness. However, normalization may cause the loss of precision, where results in:

\[
\text{normal}(g(v)) = \frac{g(v) - \min(g)}{\max(g) - \min(g)}
\]

(9)

\[
\max(\text{normal}) = 1 \quad (10) \quad \min(\text{normal}) = 0 \quad (11)
\]

For instance, in an undirected star graph, betweenness of the centre vertex would be \( (n-1)(n-2)/2 \) (1, if normalized) while betweenness of the leaves (which are contained in no shortest routes) would be 0. The examples of betweenness centrality, closeness centrality, degree centrality in the same graph is showed in figure 6.

We have another similarity matrix \( S_r \) based on degree centrality \( C_d(v) \), closeness centrality \( C_c(v) \) and betweenness centrality \( C_b(v) \), \( v \) is the vertex of topology network. The model combines dynamic origin/destination flows with the space distance data and topology network structure data between bicycle stations to construct a similarity matrix \( S \). The \( S \) can be represented as follow:

\[
S(v_i) = \alpha S_n(v_i) + (1 - \alpha) S_c(v_i)
\]

(12)

3.3. The affinity propagation (AP) algorithm

In AP algorithm, there are two kinds of message exchanged between data points responsibility and availability, which is showed in figure 7. The \( \text{res}(i, k) \), sent from data point \( i \) to candidate exemplar point \( k \) reflects the accumulated evidence for how well-suited point \( k \) is to serving as the exemplar for point \( i \), taking into account other potential exemplars for point \( i \). The \( \text{ava}(i, k) \), sent from candidate exemplar point \( k \) to point \( i \), reflects the accumulated evidence for how appropriate it would be for point \( i \) to choose point \( k \) as its exemplar, taking into account the support from other points. Bigger \( \text{res}(i, k) \) and \( \text{ava}(i, k) \) indicate that point \( k \) is more likely a centre of the cluster, and the point \( i \) is more likely to be dispatched to the cluster whose centre is point \( k \). The procedure of the AP algorithm is as follows:
First, calculate the similarity \( s(i, k) \) between \( N \) data points, construct the similarity matrix \( S \). Second, initialize the reference values, given the iteration number \( N \). Third, calculate the value of responsibility:

\[
res(i, k) = s(i, k) - \max\{ava(i, k') + s(i, k')\}, k \neq k'
\]

Forth, calculate the value of availability:

\[
ava(i, k) = \min\left\{0, res(k, k) + \sum_{i'} \max\{0, res(i', k)\}\right\}, i' \notin i, k
\]

\[
ava(k, k) = \sum_{i'} \max\{0, res(i', k)\}, i' \neq k
\]

Fifth, determine whether point \( k \) is the clustering point center point. In the end, iteration when the clustering centre point does not change in the process of continuous \( N \) iterations.

3.4. Coupling analysis between scheduling regions

Suppose there are two scheduling regions, \( i \) and \( j \). In a certain time period, \( Q(i, j) \) bikes are leased from \( i \) and finally returned to \( j \). At the same time, \( Q(j, i) \) bikes are leased from \( j \) and returned to \( i \). The value of \( Q(i, j) \), which is the sum of \( Q(i, j) \) and \( Q(j, i) \), denotes the total number of bicycles leased and returned between \( i \) and \( j \). The value reflects the correlation degree between two scheduling regions, that’s to say, the coupling degree. The larger the value \( Q(i, j) \) is, the more people use public bicycles to travel between two regions, which shows that the flow of bicycles between two regions is more frequent. So the scheduling workload between the regions is reduced by minimizing the connection between the regions, we use the coupling function \( R \) shown in Equation 9 to measure the rationality of the scheduling region partition methods:

\[
R = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} Q(i, j)}{\sum_{i=1}^{m} \sum_{j=1}^{m} r_{ij}}, \text{where } i \neq j
\]

This function \( (R) \) indicates the average coupling degree of \( m \) scheduling regions. The greater the value \( R \) is, the higher the bicycle rental frequency is between different regions, which mean that the result of scheduling region partition is more unreasonable.

4. Analysis of the experimental data and results

4.1. Experimental data acquisition

Since Hangzhou is a famous tourist city, holiday data is short-term and redundant [21]. Based on these considerations, we used 4,234,125 public bicycle data for the three-week period between November 9, 2015 and November 29, 2015. Table 1 is a partial list of the loan records during this period.

**Table 1.** Loan records of public bicycle (partial).

| Rec Id   | Bike Id | Lease Time      | Lease Station Id | Return Time     | Return Station Id |
|----------|---------|-----------------|------------------|-----------------|------------------|
| 367637055 | 684616  | 2015/11/10 13:03:04 | 2004             | 2015/11/10 13:37:17 | 4051             |
| 363712440 | 928577  | 2015/11/13 12:56:14 | 2102             | 2015/11/13 13:20:05 | 2102             |
| 363712451 | 684169  | 2015/11/15 10:55:34 | 4010             | 2015/11/15 12:01:39 | 4073             |
| 363705113 | 900969  | 2015/11/16 06:46:21 | 3461             | 2015/11/16 07:24:39 | 1085             |
| 363876543 | 903065  | 2015/11/28 06:42:01 | 4050             | 2015/11/28 06:55:01 | 2164             |
| 364394169 | 917037  | 2015/11/29 19:06:50 | 4371             | 2015/11/29 19:24:35 | 4128             |

**Table 2.** Basic data of the public bicycle stations (partial).

| Station Id | Area Id | Station Name          | Stall Num | Coordinate X | Coordinate Y |
|------------|---------|-----------------------|-----------|--------------|--------------|
| 4209       | 4       | Government Building   | 12        | 120.161603   | 30.278808    |
| 4186       | 4       | Lishui Road East      | 11        | 120.147701   | 30.33804     |
| 1403       | 1       | ZJ Second Hospital    | 21        | 120.1828     | 30.257676    |
| 5288       | 5       | YiLe Village          | 21        | 120.176611   | 30.203582    |
| 3287       | 3       | 305 Bus Station       | 21        | 120.227828   | 30.326144    |
| 2310       | 2       | Hangzhou Museum       | 21        | 120.169957   | 30.271295    |
4.2. Experimental results
We used the bike id, lease station id and return station id in table 1 to obtain the lease/return matrix \(\gamma\). Using the data in table 2. The Euclidean distance between any two stations can be obtained, and the distance matrix \(\theta\) can be built. The lease/return matrix \(\gamma\) are normalized, and so the similarity matrix \(S_n\) can be established by \(S_{ij} = -\omega_{ij} * d_{ij}\). In addition to \(S_n\), we calculate the topology network vertices similarity of bicycle stations \(S_t\) base on degree centrality, closeness centrality and betweenness centrality. The similarity of bicycle can be expressed as \(S = \alpha S_n + (1 - \alpha) S_t\). We used the AP algorithm to cluster, and obtain the partition result of the scheduling regions. Figure 8. Shows the partition result of the scheduling regions. As shown by the clustering result, the 2,222 Hangzhou bicycle stations are partitioned into 38 scheduling regions. In figure 8, a scheduling region is clearly partitioned with clear boundaries. The result is displayed on the actual road network, as is shown in figure 9 without crossing rivers, lakes, and mountains.

![Figure 8. Cluster result using our model.](image1)

![Figure 9. The result of scheduling region division on the actual road network.](image2)

4.3. Comparison with other methods
We compared our method with the existing administrative division method, the k-means algorithm, and the original affinity propagation (AP) algorithm without considering public bicycle rental. By calculating the objective function \(R\) for the different method, the results of each method are shown in table 3.

| Similarity matrix | Administrative division | K-means | AP | Our method |
|-------------------|-------------------------|---------|----|------------|
| \(S_n\)           | 0.3251%                 | 0.1462% | 0.1589% | 0.1207% |
| \(\alpha S_n + (1 - \alpha) S_t\) | 0.2834% | 0.0521% | 0.0545% | 0.0517% |

5. Conclusions
In this study, the new scheduling region partition model that achieves improved efficiency in the bicycle rentals between stations is proposed, which combines lease/return rate with space distance data and topology network structure data between public bicycle stations. The AP algorithm has been applied to the field of public bicycle research for the first time. Experimental results show that the coupling degree of the proposed model is lower than that achieved by other methods, and that the partition results are reasonable and scheduling workload is mainly concentrated in each region. Our research in the paper has great help for designing the shortest scheduling route of dispatching vehicles in each region, and can serve as a reference for future research in PBS.

Acknowledgments
This study was supported by the National Natural Science Foundation of China (No. 61602141). The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.
Reference

[1] Lathia N, Ahmed S & Capra L 2012 Measuring the impact of opening the London shared bicycle scheme to casual users *Transportation Research Part C Emerging Technologies* **22**(5) 88-102

[2] Ghosh S, Trick M & Varakantham P 2016 Robust repositioning to counter unpredictable demand in bike sharing systems *International Joint Conference on Artificial Intelligence* pp.3096-3102 AAAI Press

[3] Kloimüllner C, Papazeck P, Hu B & Raidl G R 2015 A Cluster-First Route-Second Approach for Balancing Bicycle Sharing Systems *International Conference on Computer Aided Systems Theory* pp.439-446 Springer, Cham

[4] Cruz F, Subramanian A, Bruck B P & Iori M 2016 A heuristic algorithm for a single vehicle static bike sharing rebalancing problem *Computers & Operations Research* **79** 19-33

[5] Li Y, Szeto W Y, Long J & Shui C S 2016 A multiple type bike repositioning problem *Transportation Research Part B Methodological* **90** 263-278

[6] Benchimol M, Benchimol P, Chappert B, Taille A D L, Laroche F & Meunier F et al. 2011 Balancing the stations of a self service “bike hire” system *RAIRO - Operations Research* **45**(1) 37-61

[7] Chemla D, Meunier F & Calvo R W 2013 Bike sharing systems: solving the static rebalancing problem *Discrete Optimization* **10**(2) 120-146

[8] Tal Raviv & Ofir Kolka 2013 Optimal inventory management of a bike-sharing station. *Iie Transactions* **45**(10) 1077-1093

[9] Gaspero L D, Rendl A & Urli T 2013 A Hybrid ACO+CP for Balancing Bicycle Sharing Systems *Hybrid Metaheuristics* Vol.**7919** pp198-212

[10] Shu Jia, Chou, Mabel C, Liu & Qizhang et al. 2013 Models for Effective Deployment and Redistribution of Bicycles within Public Bicycle-Sharing Systems *INFORMS*

[11] Nair R & Miller-Hooks E 2011 Fleet Management for Vehicle Sharing Operations *INFORMS*

[12] Borgnat P, Abry P, Flandrin P & Rouquier J B 2009 Studying Lyon's Vélo'V: A Statistical Cyclic Model *Complex System Society*

[13] Pierre Borgnat, Patrice Abry, Patrick Flandrin, Jean-Baptiste Rouquier & Eric Fleury 2011 shared bicycles in a city: a signal processing and data analysis perspective *Advances in Complex Systems* **14**(03) 415-438

[14] Kabra A, Belavina E & Girotra K 2015 Bike-Share Systems: Accessibility and Availability *INSEAD Working Papers Collection* (8) 1–35. https://doi.org/10.1007/978-3-319-30094-8

[15] Wen M & Kang R 2011 Some optimal models for facility location–allocation problem with random fuzzy demands *Applied Soft Computing* **11**(1) 1202-1207

[16] Solomon M M 1987 Algorithms for the vehicle routing and scheduling problems with time window constraints *Operations Research* **35**(2) 254-265

[17] Fishman E, Washington S, Haworth N & Watson A 2015 Factors influencing bike share membership: an analysis of Melbourne and Brisbane *Transportation Research Part A* **71** 17-30

[18] Zhang W, Li S & Pan G 2012 Mining the semantics of origin-destination flows using taxi traces *In Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp’ 12* p.943 NewYork, USA: ACM Press. https://doi.org/10.1145/2370216.2370425

[19] Vogel M, Hamon R, Lozenguez G, Merchez L, Abry P & Barnier J et al. 2014 From bicycle sharing system movements to users: a typology of vélo’v cyclists in lyon based on large-scale behavioural dataset *Journal of Transport Geography* **41** 280-291

[20] Bavelas A 1950 Communication patterns in task - oriented groups *Journal of the Acoustical Society of America* **22**(6) 725-730

[21] Kaplan S, Manca F, Nielsen T A S & Prato C G 2015 Intentions to use bike-sharing for holiday cycling: an application of the theory of planned behavior *Tourism Management* **47** 34-46