Latent Flow Patterns Discovery by Dockless Bike Trajectory Data for Demand Analysis

Chao Ling, JingJing Gu(✉), and Ming Sun

College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
{lingchao,gujingjing}@nuaa.edu.cn, 15651648110@163.com

Abstract. The dockless shared bikes flourish as a new concept in recent years. It allows users to find bikes anywhere via a GPS-based mobile application, and flexible cycling and parking the bikes in the same way. From the bike trajectory data produced by Users, we can extract bike flow patterns for better urban planning and Point-of-Interest (POI) recommendation. In this paper, through conducting the spatio-temporal representations of bike activity acquired from bike trajectory logs, we first design a graph clustering model With sparsity constraints that combine time information to explore potential patterns of bike flow. Next, by comparing historical trajectory logs and POI information with the flow patterns, we dig out several typical categories of bike flow patterns, which can give suggestions for further urban planning and POI recommendation. Further, our experiments via Mobike trajectory data demonstrate the effectiveness of bike flow pattern discovery.

Keywords: Shared bikes · DNN · Graph clustering

1 Introduction

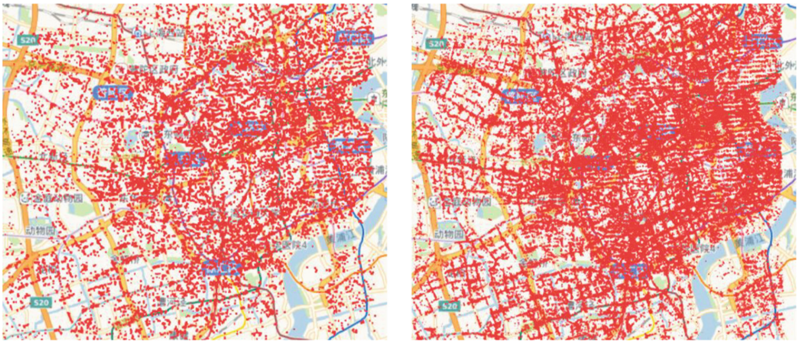
The dockless shared bikes have been receiving much attention in recent years, which change the way that people travel from motor ones to non-motor ones. Mobike[1], one of the largest bike sharing companies in the world, leads the development of bike-sharing industry and meets users' needs for more convenient short-distance travel [14]. It allows users to find, pick up and drop off their bikes anywhere through mobile applications. According to statistics from shared bikes, the orders exceed 50 million every day. What's more, we can easily analyze the temporal and spatial correlations through the trajectory data generated by the use of shared bikes, which could contribute to infer the users' preference for POI in a different time and space environments in further research. For example, riding destination could be different, such as subway stations, companies or supermarkets, and the duration of the trajectory could also change at the same

[1] https://mobike.com.

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 645–658, 2019.
https://doi.org/10.1007/978-3-030-32388-2_54

time. Therefore, exploring the latent flow patterns of shared bikes is significant to urban construction, POI recommendation, and demand analysis.

In this paper, we propose a method for extracting latent flow patterns of shared bikes trajectory data. It is still a challenging task due to the following reasons. First, the riding flow of shared bikes changes over time in a day. Figure 1 shows the geographical distribution of riding destinations at the peak and normal time respectively. Second, bike flow can also be influenced by weather, temperature, and population. Finally, the riding flow is locally-invariant and sparse, because of the short riding distance and high mobility. Previous approaches [8,15] fail to consider these factors. Therefore, how to utilize the characteristics and effectively extract the patterns of bike flow remains an open problem.



(a) The destinations in the normal time (b) The destinations in the peak time

Fig. 1. The geographical distributions at different time

To solve the problems mentioned above, we study a large set of Mobike users' spatio-temporal trajectory in Shanghai to obtain patterns of bike flow. First, we divide the whole city into small region grids and identify the primary locations by users' activities, using a density-based algorithm named *bike ordering points to identify the clustering structure* (Bike-OPTICS). Differing from traditional road segmentation and equal-size grids, this method can effectively avoid achieving undeveloped or inaccessible areas. Second, to further explore the bike usage demand in different regions, we develop a graph auto-encoder by non-linear embedding the original graphics [9]. In addition, we model the spatio-temporal interactions between region pairs with a sparsity constraint, which characterizes the locally-invariant sparse of bike flow. Finally, we run the k-means algorithm to obtain clustering result, for identifying the latent travel patterns for urban planning and Point-of-Interest (POI) demands of visitors.

Overall, the main contributions of our work are summarized as follows.

- We propose a new density-based clustering method to merge the neighboring region grids with high flow together. Based on the result of cluster analysis, we find the type attribute of each cluster.

- A deep neural network (DNN) of stacking auto-encoder with sparsity constraints is presented to identify the latent travel patterns and POI demands of visitors.
- Some suggestions about urban planning and bike migration are given through large-scale data analysis in the real world.

2 Related Work

Recently, because of the flourishing of location technology services (LTS) [4,7] and the advantage of shared bikes trajectories in improving the quality of human life, researchers are encouraged to use bike trajectory as data sources in large-scale urban user mobility studies. In the literature, existing works on bike sharing systems mainly studied the problems of further expansion of the station [11], shared bikes traffic prediction [6] and rebalance scheduling [10,12]. For example, Bao et al. [3] proposed a data-driven approach to develop bike lane construction plans based on large-scale real-world bike trajectory data. Ai et al. [1] developed a convolutional long-term memory network (conv-LSTM) method to predict the short-term spatio-temporal distribution of bikes, which reduced the space dependence and time dependence of bikes. However, these methods are not able to directly applied to dockless ones and only considered the distribution of bikes and traffic forecasts. Moreover, few studies have focused on further analysis and exploration of bike traffic patterns.

There has already been lots of research looking at extracting latent patterns [8,15]. Zhou et al. [16] proposed a topic-based model to discover latent patterns of urban cultural interactions. Gao et al. [13] discovered human lifestyle by creating a topic model from their digital footprints and social links. Ziyatdinov et al. [9] proposed a pattern extraction method for multi-view data using spectral clustering algorithm.

Currently, some studies prove that the topic model is more effective for discovering potential movement patterns [16]. However, for bike research, problems do exist with this model. First, urban data is quite sensitive to time, and these temporal flow patterns in shared bikes can hardly be captured by topic models [14]. Second, there is no corresponding assessment measure in the topic model for shared bikes flow pattern analysis. Finally, bike flow data is a graph structure, which increases the difficulty of data disposal course. Thus, we exploit a graph auto-encoder with sparsity constraints to identify the latent travel patterns, which could better reflect the structure of the graph and the spatial interaction between the pairs of regions.

3 Problem Formulation

In this study, we use two sets of real-world data collected from Mobike, including bike trajectory logs and urban POI data. Specifically, the bike trajectory logs contain the use of records from bike users in Shanghai. Table 1 shows an example of the trajectory logs. Each record consists of a bike label, pick-up time and drop-off time, and the corresponding origin and destination with detailed GPS coordinates. Besides, most of the locations can be linked to a specific POI.

Table 1. An example of the trajectory logs

Bike ID	Pick-up time	Drop-off time	Trip origin	Trip destination
8621525316	1517511153	1517511619	121.45,31.20	121.43,31.22
8621633865	1517501153	1517501779	121.44,31.21	121.43,31.22
8621399390	1517511233	1517521779	121.43,31.25	121.43,31.24
8621399332	1516521133	1517123779	121.37,31.21	121.40,31.25
8621399312	1517521133	1517523779	121.40,31.21	121.42,31.25

This paper has two tasks. Specifically, (1) region partition and flow matrix construction, and (2) flow pattern extraction. In the first task, we aim to discover the integrated urban areas with high flow density and construct a flow matrix for the trajectory logs based on the travel flow information extracted from shared bikes. The ultimate goal of the second task is to extract flow patterns, by which we can perform urban planning and POI recommendations.

The investigation framework of our work is presented in Fig. 2. We firstly conduct the data preprocessing, and divide the urban area into small area grids. Secondly, by extracting the travel flows in these grids, we construct a flow matrix. Next, the similarity matrix of the flow matrix is constructed based on the similarity measurement. Finally, stacking auto-encoder is applied to the original graphics. This model can learn users' spatial-temporal preferences by studying their behaviors during the process of cycling, which can supply the travel demand analysis and targeted POI recommendation with strong supports.

4 Methodology

4.1 Region Partition and Flow Matrix Construction

The dockless shared bikes are different from the traditional station-based bikes due to their flexibility. The former doesn't need to be parked and locked at the designated places, which presents a challenge to the research. Therefore, constructing flow adjacency matrix is not easy. However, we find that the riders often come together automatically, when they have similar destinations. Inspired by this observation, we decide to cluster bike parking points into regions. Through the clustering results, we can make the flow adjacency matrix.

In this work, we find that the OPTICS algorithm [2] is a suitable approach for our problem of region partition. This method requires two parameters as input: the maximum radius *eps* for searching, and the least number of points *minpts* to form a cluster. The parking positions of bikes vary greatly, which needs to set the bike maximum radius carefully. With this limitation in mind, as Algorithm 1

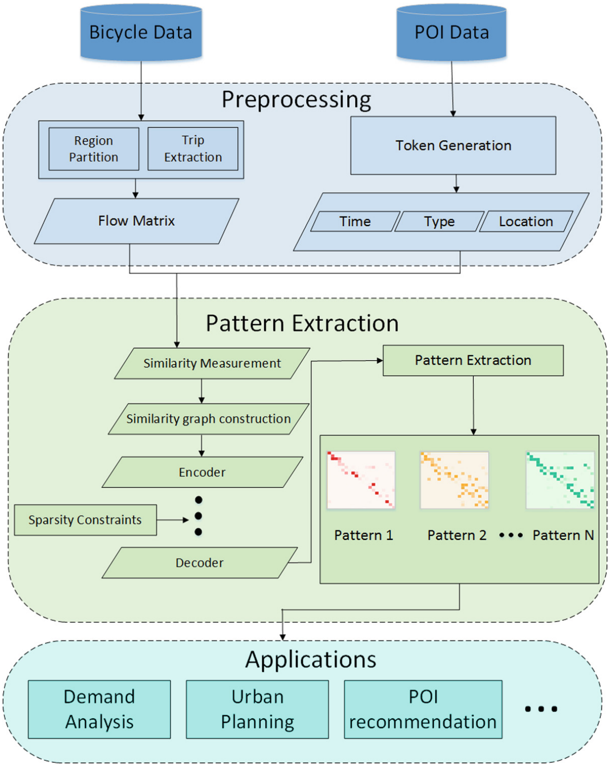
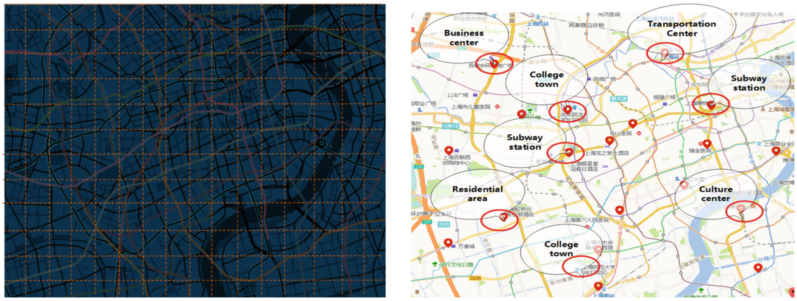


Fig. 2. Framework overview

shows, we propose a modified version named Bike-OPTICS that defines different reachable distances for different areas. In Bike-OPTICS, we collect all drop-off positions. In the procedure of clustering, the areas where bikes are densely distributed should be given higher weights. We define the maximum radius as:

$$MR = eps - \epsilon * \frac{getNeighbors(p, eps)}{N}, \quad (1)$$

where ϵ is set to be a small constant such as 0.01. Generally, the urban areas are split into small grids at first, as shown in Fig. 3(a). Then we start clustering using one of the grids, which is never categorized into existing clusters. Next, the rest grids which are reachable to the current grid to these new clusters are added. This process is repeated until all the grids are clustered, and no new cluster is created.



(a) Grid-based partition (b) Some landmarks

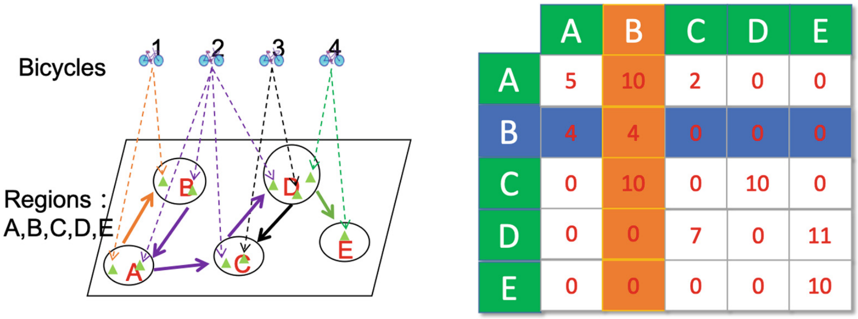
Fig. 3. The urban area notations of Shanghai

Algorithm 1. Bike-OPTICS

```
Input:
    $DB = [...(p_i, location)...]$, maximum radius $eps$.
Output:
    center point of clusters $order = [c_1, c_2, ...c_n]$, cluster groups of points $c_{points} = [L_1^*, L_2^*...L_n^*]$, $L_i^* = [p_1, p_2...p_n]$
initialize $Order = list(), Se = list(), RD = list(maxdis), RD(0) = 0, MR = list(...MR_i...)$;
for each unprocessed point $p$ of $DB$ do
    $MR \leftarrow getMR(p, eps)$;
    mark $p$ as processed;
    $Order \leftarrow p$;
    update RD;
    $Se \leftarrow getNeighbors(p, MR)$;
    for each unprocessed $q$ in $Se$ do
        $MR \leftarrow getMR(q, eps)$;
        marked q as processed;
        $Order \leftarrow q$;
        update RD;
        $Se \leftarrow getNeighbors(q, MR)$;
    end for
end for
for each $p$ in $DB$ do
    compute Reachable Distance to each point $c_i$ in Order;
    if Reachable Distance $< MR_i$ then
        $L_i \leftarrow p$;
    end if
end for
```

Our algorithm is similar but different from the classic Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [5]. In the DBSCAN algorithm, a radius and a threshold should be defined beforehand, but for Bike-OPTICS algorithm, it generates an augmented cluster ordering for cluster analysis, rather than clustered results explicitly. It reflects all the results from density-based clustering in any parameters setting. In other words, clustering based on any radius and any threshold can be derived from this ordering. For the sake of demonstration, we draw some landmarks in Fig. 3(b).

With the partitioned regions, we can construct flow matrix and flow tensor. After partitioning the entire city, a bike flow matrix that records flow between any two regions in the same time segment can be constructed. Given a whole city divided into M regions, the flow matrix is defined as $F^t = \{f^t_{ij}\} \in R^{M \times M}\}$, where f^t_{ij} denotes the number of rides from the ith region to the jth region in the t time fragment.



	A	B	C	D	E
A	5	10	2	0	0
B	4	4	0	0	0
C	0	10	0	10	0
D	0	0	7	0	11
E	0	0	0	0	10

Fig. 4. A example of bike flow matrices construction

Figure 4 shows an example of the bike flow matrix structure. The x-value and y-value represent the starting region and the ending region of one bike trajectory respectively, and each matrix cell represents the bike flow between the two regions in a time slice. Particularly, we can also get the total pick-up flow or drop-off flow of region i (f^p_i and f^d_i) by calculating the sum of the corresponding column or a row. We have N flow matrices through which we build flow tensors $\mathscr{F} = \{F^{t_1}, F^{t_2}, \ldots, F^{t_N}\}$. In addition, we set the flow matrix interval as one hour.

4.2 Flow Pattern Extraction

In this section, we present a new graph clustering model based on the auto-encoder with sparsity constraints for extraction latent flow pattern. Due to the property of the bike trajectory data, the structure of the flow matrix is non-Euclidean, which does not have translation invariance. The traditional cluster methods cannot extract its internal structure very well. Therefore, we apply the deep learning method of graph clustering. Moreover, we add the sparsity constraints due to the local invariance of the bike flow.

We use an auto-encoder based on graph clustering model, which is a key component of a deep neural network. Figure 5 shows the main architecture of the auto-encoder. In order to rectify the problem of overfitting, we integrate the dropout layer into the DNN model.

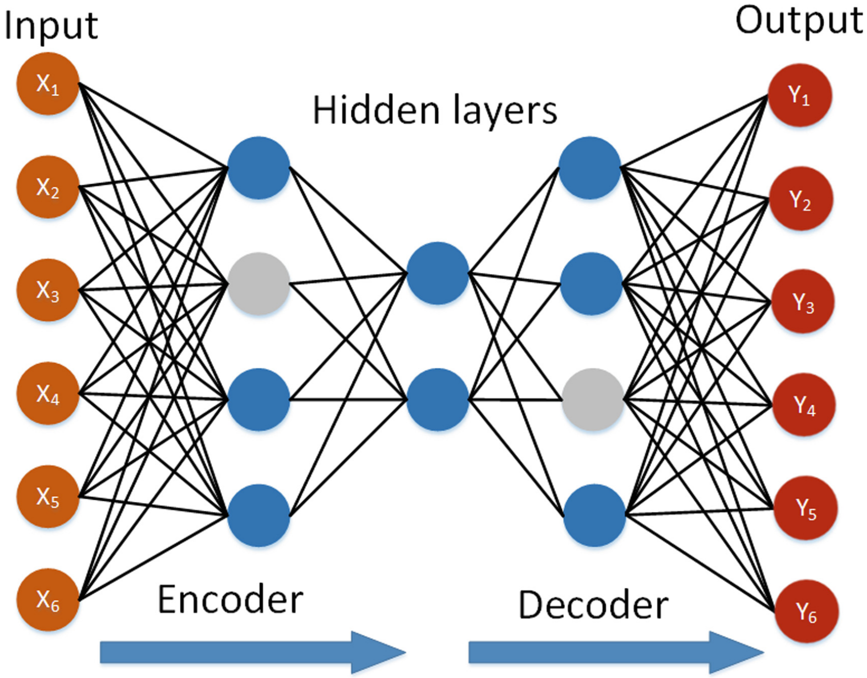


Fig. 5. DNN architecture

As stated in the previous section, an n-node graph G can be represented by its similarity matrix S. We use the Radial Basis Function (RBF) function to define the similarity matrix, which is widely used in the field of the similarity matrix:

$$S_{ij} = \exp(-\frac{\|F^i - F^j\|_2^2}{2\sigma^2}) \tag{2}$$

Next, we normalize the training set in the DNN and use the output features of the deepest layer as the graph embedding. Finally, we utilize the k-means algorithm for producing the final clustering result of graph embedding.

```
Algorithm 2. Clustering with sparse GraphEncoder
Input:
    Graph $G$, flow tensor $\mathscr{F}$, similarity matrix S, degree matrix $D$, DNN layer numbers
    $\ell$, Dropout layer parameter $p$
Output:
    Clustering result.
initialization $a^0 = D^{-1}S$
for $i = 0$ to $\ell$ do
    Build a DNN architecture with input $a^{(i)}$;
    Multiply by the Dropout layer $r^{(l)}$,$a^{(i)} = r^{(i)} * a^{(i)}$;
    Train the DNN by optimizing (7) with sub-gradient method, Obtain the hidden
    layer activations $h^{(i)}$;
    Let $a^{(i+1)} = h^{(i)}$ ;
end for
Run k-means on $a^{\ell} \in R^{n \times n^{(\ell)}}$;
```

Specifically, for a DNN model with L layers, the output of a hidden layer can be described as:

$$r^{(l)} \sim Bernoulli(p), \tag{3}$$

$$\tilde{y}^{(l)} = r^{(l)} * y^{(l)}, \tag{4}$$

$$z^{(l+1)} = w^{(l+1)}\tilde{y}^{(l)} + b^{(l+1)}, \tag{5}$$

$$y^{(l+1)} = f(z^{(l+1)}). \tag{6}$$

where $l = 1 \ldots L$, $a^0 = D^{-1}S$ is the input layer with feature vector x. w^l and b^l denote the DNN weight matrix and bias of the l-th hidden layer. $f(\cdot)$ represents the non-linear activation function of the hidden layer and the output layer, such as ReLu function:

$$ReLu(z) = max(0, z) \tag{7}$$

and tanh function:

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{8}$$

Here, D is the diagonal matrix with the node degrees in the corresponding diagonal elements, and Q is the normalized Laplacian matrix. Due to the property of the Laplacian matrix, Q is symmetric. Then, the optimization goal is to minimize the reconstruction error between the original data x and the reconstructed data y.

$$\arg\min_{\theta \in \Theta} -\frac{1}{n}\sum_{i=1}^{n}[x \ln y - (1-x)\ln(1-y)]. \tag{9}$$

We also impose the sparsity constraints to the activation in the hidden layer, the loss function is:

$$Loss(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[x \ln y - (1-x)\ln(1-y)] + \beta\|\hat{a}\|_1, \tag{10}$$

where β controls the weight of the sparsity penalty, and $\hat{a} = \frac{1}{n}\sum_{j=1}^{n} h_j$ is the average of the hidden layer activations.

For the Eq. 7, we can use back-propagation for training. Since the l_1-norm is non-differentiable and cannot be solved with the traditional gradient descent, we use the sub-gradient method to solve it.

Using the output of the hidden layer as the input to the next layer, we use the intermediate output of the encoder as a new representation of graph and run k-means on it to get the clustering results.

5 Evaluation

In this section, we first introduce the settings of the experiment including dataset, baseline algorithms and evaluation criteria that we use in the course of the experiment. Then we show the experimental results and give an in-depth analysis, which proves the superiority of our algorithm.

5.1 Experiment Settings

Dataset. The dataset we used covers 177, 357, 367 riding records, generated by 389, 703 shared bikes ranging from December 2017 to July 2018 in Shanghai city. Each record consists of a bike label, the pick-up time and drop-off time, as well as the corresponding origin and destination with detailed GPS coordinates.

Evaluation Criteria. Davis-Bolding Index (DBI), also known as the classification suitability index, is the sum of the average distance $avg(C)$ between the two clusters C_i and C_j divided by the distance between their center points u. The smaller the DBI, the better the clustering effect. Let $C = \{C_1, C_2, \ldots, C_k\}$,

$$avg(C) = \frac{2}{|C|(|C|-1)} \sum_{1 \le i < j \le |C|} dist(x_i, x_j), \tag{11}$$

$$u = \frac{1}{|C|} \sum_{1 \le i \le |C|} x_i, \tag{12}$$

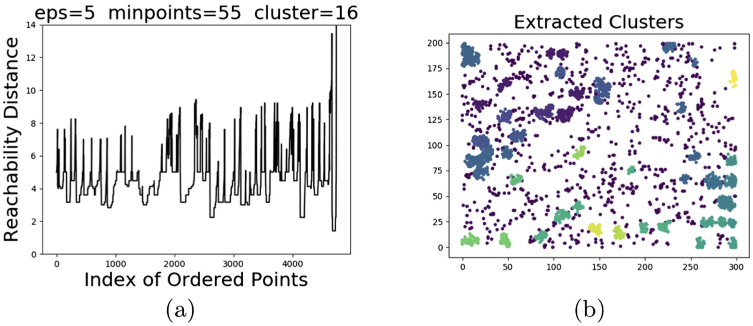
$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \ne j} \left(\frac{avg(C_i) + avg(C_j}{dist(u_i, u_j)} \right). \tag{13}$$

Baselines. We use the following methods as baseline algorithms.

(1) Spectral Clustering. Spectral clustering is an algorithm that evolves from the graph theory and has been widely used in clustering. Compared with the traditional k-means algorithm, spectral clustering is more adaptable to our data distribution, and has better clustering results.
(2) K-means. In order to verify the validity of our method, we perform the k-means algorithm on the original graph structure.

5.2 Experimental Results

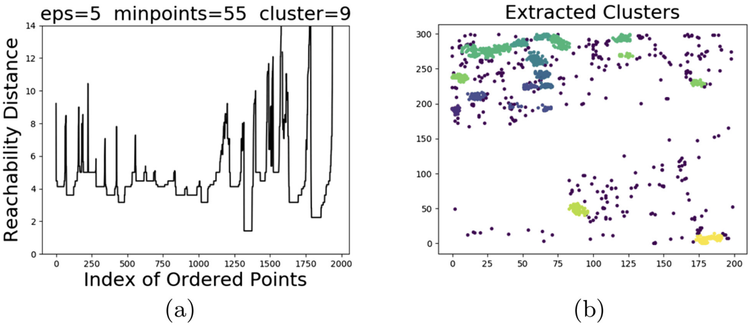
We selected two different hot spot regions to display our experimental results. At first, we performed the Bike-OPTICS algorithm using different parameters on these regions.



(a)

(b)

Fig. 6. Region partition in Grand Theatre

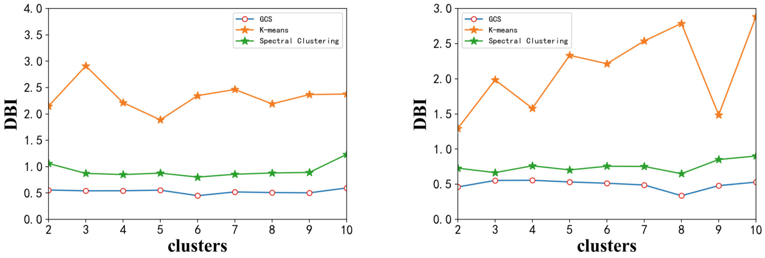


(a)

(b)

Fig. 7. Region partition in Grand Theatre

Figure 6 shows the results of the Bike-OPTICS algorithm in different regions. We choose Shanghai Grand Theatre and Shanghai World Expo Hall as center respectively, and take their surrounding areas within 2000 m radius as research regions. From Figs. 6 and 7, we can see that the two regions are divided into different numbers of clusters in Figs. 6(a) and 7(a). And in Figs. 6(b) and 7(b), every valley represents a cluster. Among them, purple represents noise which does not form any clustering. The Shanghai Grand Theatre locates at the center of Shanghai and the bikes distribute densely there, which results in more clusters. In contrast, the Expo Park is located in the suburbs of Shanghai, the bike distribution is very scattered and cannot form effective clusters.



(a) Grand Theatre (b) Expo Park

Fig. 8. Clustering results

The experimental results of the three algorithms above are shown in Fig. 8, with a horizontal axis representing the number of the predefined clusters, and the vertical axis representing the corresponding DBI value. We can see that: (i) Graph cluster with sparse constraint (GCS) outperforms the spectral clustering, which indicates that graph clustering with sparse constraint helps to improve the effectiveness of clustering. (ii) K-means can't handle graph structure very well, as the DBI value of k-means is much higher than spectral clustering and GCS. (iii) Different regions have different clustering results. In Fig. 8(a), GCS has the lowest DBI value when cluster number is 6, while it has the lowest DBI when clustering is 9 in Fig. 8(b), possibly because Expo Park is located in the suburbs of Shanghai and the activities are more diverse there.

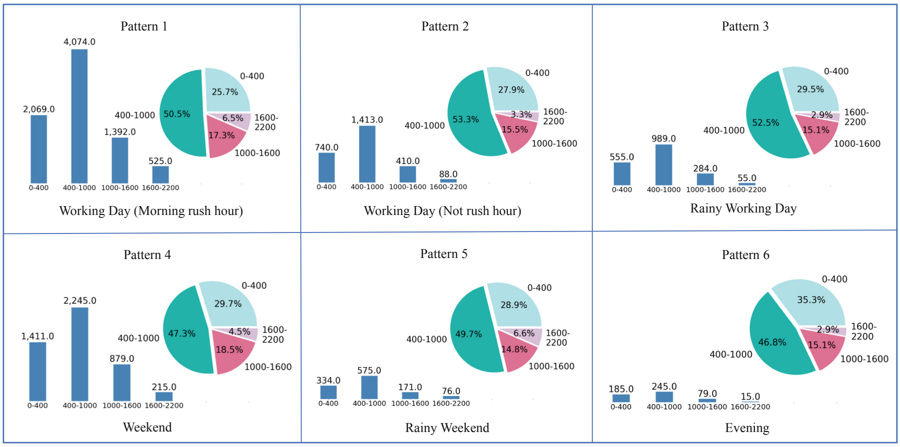


Fig. 9. Region partition in Grand Theatre

To further evaluate the performance of the model we proposed, we visualize and analyze the experimental results in another way. As shown in Fig. 9, our method has obtained six bike flow patterns including working day (morning rush hour), rainy working day and weekend, etc, regarding Shanghai Grand Theatre area as a study area. This proves that our method is effective for extracting bike flow patterns.

6 Conclusion

In this paper, we introduced a new method on bike flow patterns analysis through the bike trajectories data and POI data. At first, we divided urban Shanghai into small grid areas, and then proposed a density-based clustering method for merging neighboring grid areas into a cluster. With the clustering results, we constructed the bike flow matrix, which recorded flow between any two regions in the same time segment. To further explore the users' behavior patterns, we developed a graph clustering model with sparsity constraints. Finally, we estimated the performance of the model based on the large-scale real-world data collected from Mobike in Shanghai. The experimental results show that the method we proposed can effectively extract the bike flow pattern. In future work, we will apply the presented method to the demand analysis, POI recommendation, and other tasks.

References

1. Ai, Y., et al.: A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system. Neural Comput. Appl. **31**, 1–13 (2018)
2. Ankerst, M., Breunig, M., Kriegel, H., Ng, R., Sander, J.: Ordering points to identify the clustering structure. In: Proceedings of ACM SIGMOD, vol. 99 (2008)
3. Bao, J., He, T., Ruan, S., Li, Y., Zheng, Y.: Planning bike lanes based on sharing-bikes' trajectories, pp. 1377–1386 (2017)
4. Bauer, S., Noulas, A., Seaghdha, D.O., Clark, S., Mascolo, C.: Talking places: modelling and analysing linguistic content in foursquare, pp. 348–357 (2012)
5. Ester, M., Kriegel, H.P., Xu, X.: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise (1996)
6. Gast, N., Massonnet, G., Reijsbergen, D., Tribastone, M.: Probabilistic forecasts of bike-sharing systems for journey planning, pp. 703–712 (2015)
7. Hasan, S., Ukkusuri, S.V.: Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C-Emerg. Technol. **44**, 363–381 (2014)
8. Hong, L., Zheng, Y., Yung, D., Shang, J., Zou, L.: Detecting urban black holes based on human mobility data. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 35. ACM (2015)
9. Kanaanizquierdo, S., Ziyatdinov, A., Pereralluna, A.: Multiview and multifeature spectral clustering using common eigenvectors. Pattern Recogn. Lett. **102**, 30–36 (2018)
10. Liu, J., Sun, L., Chen, W., Xiong, H.: Rebalancing bike sharing systems: a multi-source data smart optimization, pp. 1005–1014 (2016)

11. Liu, Z., Shen, Y., Zhu, Y.: Where will dockless shared bikes be stacked?:—parking hotspots detection in a new city. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 566–575. ACM (2018)
12. Pan, L., Cai, Q., Fang, Z., Tang, P., Huang, L.: Rebalancing dockless bike sharing systems. CoRR abs/1802.04592 (2018). http://arxiv.org/abs/1802.04592
13. Tian, F., Gao, B., Cui, Q., Chen, E., Liu, T.: Learning deep representations for graph clustering, pp. 1293–1299 (2014)
14. Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., Moscibroda, T.: Mobility modeling and prediction in bike-sharing systems, pp. 165–178 (2016)
15. Yuan, N.J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., Xiong, H.: Discovering urban functional zones using latent activity trajectories. IEEE Trans. Knowl. Data Eng. **27**(3), 712–725 (2015)
16. Zhou, X., Noulas, A., Mascolo, C., Zhao, Z.: Discovering latent patterns of urban cultural interactions in WeChat for modern city planning. In: Knowledge Discovery and Data Mining, pp. 1069–1078 (2018)