Proceedings of the 4th World Congress on Civil, Structural, and Environmental Engineering (CSEE'19)
Rome, Italy – April 7-9, 2019
Paper No. ICEPTP 149
DOI: 10.11159/iceptp19.149

Controlling the blind expansion of dockless shared bicycles through travel demand identification and prediction

Jian Wang[1], Yuan Tian[1], Siqing Wang[1]
[1]School of Transportation Science and Engineering, Harbin Institute of Technology
73 Huanghe Road, Harbin, China
wang_jian@hit.edu.cn; 14b932014@hit.edu.cn; annie_wsq@163.com

Abstract **-** Free-floating bike sharing (FFBS) is an innovative bike sharing mode that is expanding rapidly around the world nowadays. FFBS has always been known to significantly reduce traffic congestion, urban traffic carbon emissions, noise, and support a greener sustainable development of urban environments. However, without knowing the users' travel demand locations and quantity, blindly expanding shared bicycles on the market has already caused shared bicycle "disasters" which affect city environment and cause serious waste of resources in some Chinese cities. This paper proposes a mid-long term FBBS travel demand identification and prediction methodology which is an important support for bike-sharing companies' in determining the number of shared bicycles' deployment and avoid blind expansion. The methodology first uses a spatial-temporal trajectory clustering algorithm based on an innovative spatial-temporal distance function. By clustering individual users' long-term travel trajectories, the regular travel demand and random demand can be identified. Then, the total demand to be known of a future period is determined by the sum of two parts, total regular demand based on aggregated individual regularities and total random demand predicted based on a proposed MLP Neural Net Model. At last, the proposed methodology is tested using a two weeks' dataset of shared bicycle travel trajectories from Beijing Mobike system. The results indicate that the proposed methodology based on regular travel identification and prediction has strong practicability and outperforms the method directly using total demand to predict in terms of accuracy.

Keywords**:** Free-floating bike sharing; Spatial-temporal trajectory clustering; Travel demand identification; Demand prediction; Sustainable development.

1. Introduction

Nowadays, promoting cycling is one of the most popular policies implemented by local governments to encourage greener mobility and make cities more liveable. In this context, bike-sharing systems are growing rapidly around the world. With the rapid development of the internet and smartphone technology, mobile payment has become increasingly prevalent. Consequently, free-floating bike sharing (FFBS) systems have been widely adopted in major cities across mainland China [1]. FFBS is an innovative bike sharing mode that provide fully dockless services.

As serving 12 countries and more than 200 cities around the world, an average of 30 million rides per day, over 18.2 billion kilometres accumulative total travel distance, reducing carbon emissions by 4.4 million tons in around 20 months, one Chinese FFBS Enterprise Mobike won the "Earth Guardian" award from the United Nations Environment Programme in December 2017, the highest award in the field of environmental protection in the world. The FFBS can significantly reduce traffic congestion, air pollution, noise and support a greener growth of urban environments.

However, some cities have suffered from shared bicycle disasters as blind expansion. Without knowing the users' travel demand locations and quantity, blindly putting shared bicycles on the market will lead to excessive shared bicycles with very low utilization in some areas, which take up a lot of public space and block the normal travel of citizens, affect the city environment and management, becoming "city garbage". Nevertheless, in some areas where there does have travel demand, but the deployment is not enough, which makes it difficult to find shared bicycles and influence users' travel. Besides, as the high cost of operation and maintenance, FFBS enterprises prefer to buy new bicycles rather than recycle them. A large number of poorly managed bicycles have become environmental burden as figure 1 shows. By June 2018, China's major shared bicycle companies have put in more than 27 million bicycles worldwide and will produce nearly 400,000 tons of scrap metal when they are scrapped, which will cause serious waste of resources and damage the environment.

This paper proposes a mid-long term FBBS travel demand prediction methodology that can assist bike-sharing companies in determining the number of deployment and avoid blind expansion of shared bicycles. The methodology first use a spatial-temporal trajectory clustering algorithm to cluster individual users' long-term travel trajectories to identify the regular travel demand and random demand. Then, total regular demand will be identified from obtained individual users' travel regularities and a MLP Neural Net Model is proposed to predict the total random demand of the future period to be predicted. At last, future total demand is determined by calculating the sum by adding identified total regular demand and predicted total random demand together. Finally, a dataset from Beijing FBBS system will verify the practicability and accuracy of proposed methodology.

The remainder of this article is organized as follows. In the next section related research are reviewed. Then Section 3 introduces the integrated methodology which can predict future total travel demand based on identified regular demand and predicted random demand. Section 4 uses the dataset of Mobike trip data from Beijing for methodology validation and presents the results. Section 5 provides the conclusions drawn in this study and suggests future research directions.

Fig. 1: Large number of poorly managed bicycles have become burdens to Chinese cities.

2. Literature Review of bike-sharing demand related studies

The rich data generated by bike-sharing system has given an important way to academics and practitioners to analyze travel demand and behaviour. For traditional station-based bike sharing (SBBS), Pierre Borgnat et al. analysed demand evolution, daily patterns and trip distributions using real data of SBBS in Lyon. Besides, trip purposes are assumed based on spatial-temporal distribution among each group of rentals [2]. Jonathan Corcoran et al. used a novel spatial analytical techniques to study the impact of weather conditions and calendar events on the Brisbane's SBBS spatial-temporal dynamics [3]. For new FBBS, Chengcheng Xu et al. developed a dynamic demand forecasting model for FBBS using the deep learning approach. Through the spatial-temporal analysis, the results indicate the imbalance of travel demand of FBBS. They developed long short-term memory neural networks (LSTM NN) to predict the bike sharing trip production and attraction at traffic analysis zones (TAZ) for different time intervals, including the 10-min, 15-min, 20-min and 30-min intervals. The experiment results validated the developed LSTM NN can be used to predict the gap between sharing bike inflow and outflow [4]. Aritra Pal et al. studied the mobility patterns and imbalance of an FFBS by analysing its historical trip and weather data. They proposed a simple method to decom-pose continuous variables into binary variables and two stage models that consider interactions between independent variables and improves the (quasi-)Poisson regression model [5]. Yu Shen et al. adopted spatial autoregressive models to analyse a dataset of GPS data of one Singapore FFBS operators for spatial-temporal travel patterns of bike usage in Singapore [6]. Maria Bordagaray proposed a data mining algorithm to analyze the bike usage casuistry within a sharing scheme. The proposed algorithm is a powerful tool to characterize the actual demand for bike-sharing systems [7]. Leonardo Caggiania et al. proposed a comprehensive dynamic bike redistribution method. The method starts from the prediction of the number and position of bikes over a system operating area and ends with a relocation Decision Support System.

The method solves the imbalance of bicycles between zones owing to one-way trips. A test case study with a detailed sensitivity analysis showed the effectiveness of the suggested method [8].

However, the existing related literatures that studied the mobility patterns and short-term travel demand prediction at traffic analysis zones are aiming at solving bike demand imbalance problem and redistribution. Research that related to long-term travel demand identification and prediction for controlling deployment and avoiding blind expansion is very limited.

3. Methodology

3.1. Spatial-temporal distance function

Spatial-temporal distance function considers spatial distance, direction distance and temporal distance here is to calculate how far apart trajectories are. Reasonable distance function will be the premise of successful clustering. Figure 2 is an illustration of spatial-temoporal distance between trajectories. S_i and E_i are starting point and ending point of Line segement L_i, S_j, E_j are starting point and ending point of Line segement L_j. P_{Si} and P_{Ei} are the projection points from Si and Ei onto L_j, respectively. t_{Si}, t_{Ei}, t_{Sj}, t_{Ej} are corresponding timestamp to strating points and ending points. θ is the included angle between L_i and L_j.

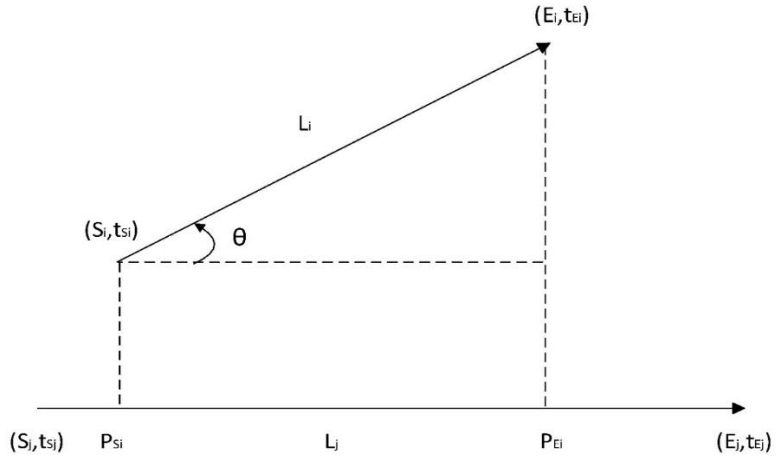

Fig. 2: Illustration of spatial-temporal distance between trajectories.

Spatial distance (SpaDist) is to measure the geographic distance difference between trajectories. Here, we followed the formula in [2].The spatial distance measure formula is as below:

$$SpaDist(L_i, L_j) = \frac{D(S_{i,}P_{Si})^2 + D(E_{i,}P_{Ei})^2}{D(S_{i,}P_{Si}) + D(E_{i,}P_{Ei})} \tag{1}$$

Where $D(S_{i,}P_{Si})$ and $D(E_{i,}P_{Ei})$ are the Euclidean distance between S_i and P_{Si}, and the distance between E_i and P_{Ei}, respectively. Directional distance (DirDist) function as below is to measure the directional difference between two line segments.

$$DirDist\left(L_i, L_j\right) = \begin{cases} \sin\theta & (0^\circ \le \theta \le 90^\circ) \\ 1 + \sin(\theta - \frac{\pi}{2}) & (90^\circ < \theta \le 180^\circ) \end{cases} \tag{2}$$

Time distance (TimeDist) is to measure the time difference between trajectory segments as shown below. Time mentioned here is unrelated to date, and just takes 24h as a cycle.

$$TimeDist\left(L_i, L_j\right) = \begin{cases} \left| \frac{t_{Sj} + t_{Ej}}{2} - \frac{t_{Si} + t_{Ei}}{2} \right| & \left(\text{if } \left| \frac{t_{Sj} + t_{Ej}}{2} - \frac{t_{Si} + t_{Ei}}{2} \right| \leq 12 \right) \\ 24 - \left| \frac{t_{Sj} + t_{Ej}}{2} - \frac{t_{Si} + t_{Ei}}{2} \right| & \left(\text{if } 12 < \left| \frac{t_{Sj} + t_{Ej}}{2} - \frac{t_{Si} + t_{Ei}}{2} \right| \leq 24 \right) \end{cases} \quad (3)$$

At last, the min-max normalisation to scale each distance measure into the unit range [0,1] is used to eliminate the scaling problem of above three measures. Besides, we set a SpaDist threshold of 500 based on considering the practical significance. If the SpaDist is greater than 500, the normalized SpaDist value will be set as two. The Spatial-temporal distance (Spa-timeDist) can be computed as below:

$$Spa-timeDist\left(L_i, L_j\right) = W_1 \times SpaDist\left(L_i, L_j\right)^* + W_2 \times DirDist\left(L_i, L_j\right)^* + W_3 \times TimeDist\left(L_i, L_j\right)^* \quad (4)$$

Where $SpaDist(L_i,L_j)^*$, $DirDist(L_i,L_j)^*$ and $TimeDist(L_i,L_j)^*$ are min-max normalised distances correponded to $SpaDist(L_i,L_j)$, $DirDist(L_i,L_j)$ and $TimeDis\ (L_i,L_j)$. W_1, W_2 and W_3 are relative weights for $SpaDist(L_i,L_j)^*$, $DirDist(L_i,L_j)^*$ and $TimeDist(L_i,L_j)^*$. In this paper, we use the same equal weight for these different distances.

3.2. Spatial-temporal trajectory clustering based on Two Step algorithm

The aim of clustering analysis of individual user-based long-term travel trajectory data is to identify the regular travel users and their regular travel demand, random users and random trips.

Two Step clustering algorithm belongs to Hierarchical Algorithms. The first step is to compress the original input data into sub-clustering which is convenient for processing. The second step is to merge small sub-clusters into larger and larger clusters step by step through hierarchical clustering. The algorithm does not need to define the number of clusters in advance, but automatically obtains the best clustering scheme by evaluating the generated clustering results. Xueling Wu et al. [9] applied Two step clustering algorithm to classify the deformation states of two typical colluvial landslides in the Three Gorges, China. The results validated the effectiveness of the Two Step algorithm.

The first step is pre-clustering stage, the clustering feature tree (CF tree) algorithm is used here. Firstly, the trajectory data in the dataset are read one by one and inserted into the CF tree to achieve the growth of CF tree. When the growth of CF tree exceeds the threshold volume, the possible outlier trajectory of CF tree is removed first, then the spatial threshold is increased and the CF tree is reduced. Then the outlier trajectory of CF tree without reduction is inserted into the CF tree. After traversing the data, the real outlier trajectory is the potential outlier trajectory that cannot be inserted into the CF tree. Finally, the clustering features of the final CF leaf element corresponding to the sub-cluster are output to the second step of the algorithm.

According to the clustering characteristics of each trajectory sub-cluster of the final leaf element of CF tree obtained in the pre-clustering stage, the trajectory sub-cluster $C_T=\{C_{T1}, C_{T2},... C_{Tn}\}$ is clustered twice. The agglomerative hierarchical clustering method is used in the clustering stage. This method recursively merges the nearest trajectory clusters to cluster. Searching for the two sub-clusters with closest spatial-temporal distance from n trajectory sub-clusters C_{T1}, C_{T2},... C_{Tn} and merge them into a new cluster. Continue searching for the nearest clusters in n-1 clusters and merge them and repeat the process until all clusters are merged into one large cluster.In the process of clustering, the first n clusters are merged into the last one. If the expected clustering results are s ($1 \leq s \leq n$), the output is the clustering results of the remaining s clusters in the clustering process.

3.3. Travel demand prediction based on MLP Neural Net Model

By clustering each individual user's travel spatial-temporal trajectories in section 2.2, the users who have regular travel demand can be identified. By collecting all the individual users' regular demand and distinguished random demand, an area's total travel demand can be divided into two parts, total regular demand and total random demand. The prediction algorithm in this paper assumes that the regular travels will recur regularly over a period of time, that is, during the period to be predicted, the users who have regular demand will use the shared bicycles according to the identified travel patterns.

For the total random travel demand of an area, a MLP Neural Net Model will be used for prediction in this paper. MLP Neural Net Model is a class of feedforward artificial neural network. An MLP consists of, at least, three layers of nodes: one input layer, one or more hidden layers and one output layer [10]. MLP can be seen as a directed graph, consisting of multiple node layers, each layer connected to the next layer. In addition to input nodes, each node is a neuron (or processing unit) with a non-linear activation function. A supervised learning method called back propagation algorithm is often used to train MLP. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. When the input is applied to the network, the network solution is compared with the target solution designed for the network, and then the learning error is calculated and used to adjust the network parameters [11].

Due to the limited period of experimental data, this paper proposes a model to predict travel demand of a day. The created network has four input variable parameters such as weather condition and day of week. And the output variable parameter is the bike-sharing travel demand of an area. Therefore, the neural network is designed with four neurons in the input layer and 1 neuron in the output layer as shown in figure 3. The MLP consists of four layers (an input and an output layer with two hidden layers) of nonlinearly-activating nodes. If data permits, adjusting the input variable parameters of the model, travel demand of week, month, quarter and even year can also be predicted.

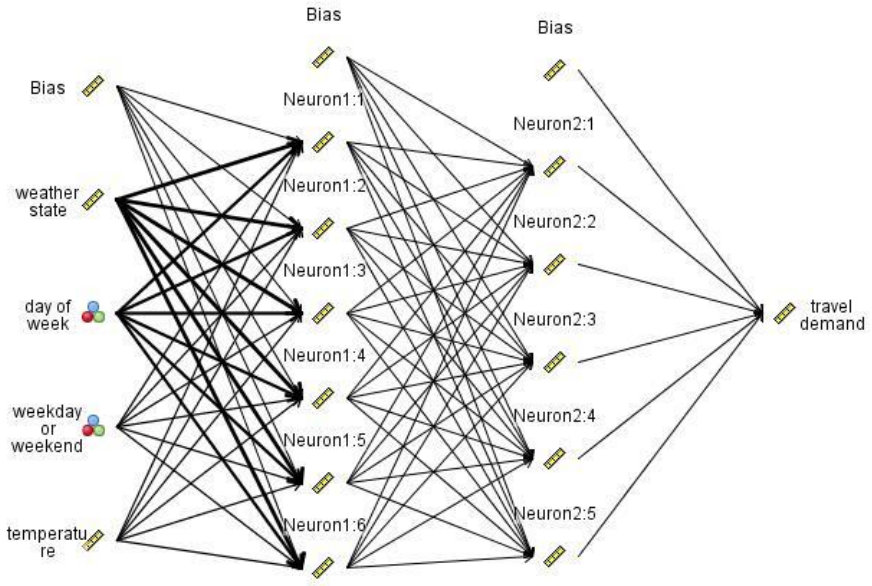

Fig. 3: A MLP neural net model to predict random travel demand constructed in this paper.

The final total travel demand of a certain area in the future will be equal to the identified regular travel demand plus the predicted random demand.

4. Application to Mobike FFBS

4.1. Data Description

A dataset from Beijing Mobike system is used to validate the methodology. It is very popular to use FFBS in Beijing, no matter residents or travelers, no matter commuting travel or casual travel. The dataset contains of 349,693 users' 3,210,496 Mobike trips from May 10 to May 16, and from May 18, to May 24, 2017. The data fields contain User ID, Order ID, Bike ID, Biketype, Start time, Start location and End location of a given record. In this paper, 10 areas on map are selected from Beijing Fengtai District, and all trips whose starting location coordinates fall within 10 areas are selected from Beijing Mobike dataset for experiment research.

4.2. Results

As the prediction target of the experiment in this paper is the travel demand of a day, May 19 and May 23 which are sunny and rainy days respectively are selected for predicting. The model uses data from May 10 to May 18 for May 19's prediction and uses data from May 10 to May 22 for May 23's prediction.

Based on spatial-temporal trajectory clustering algorithm descripted in section 2.2, all individual bike-sharing users involved in 10 areas were clustered. Figure 4 shows some of the individual users' spatial-temporal trajectory clustering results based on dataset from May 10 to May 22. The three coordinates in figure 4 are longitude, latitude and time, respectively. The arrows indicate the direction of the trips. The red directed line segments are outlier trajectories and the other colours represent the clusters. The users of ID 3967 and ID 6819 have both regular travel trajectory clusters and outlier trajectories. The regular travel trajectory clusters are identified as regular demand. The red outlier trajectories are identified as random demand. The users of ID 8087 and 1247 are very regular travel users who do not have random travel trips.

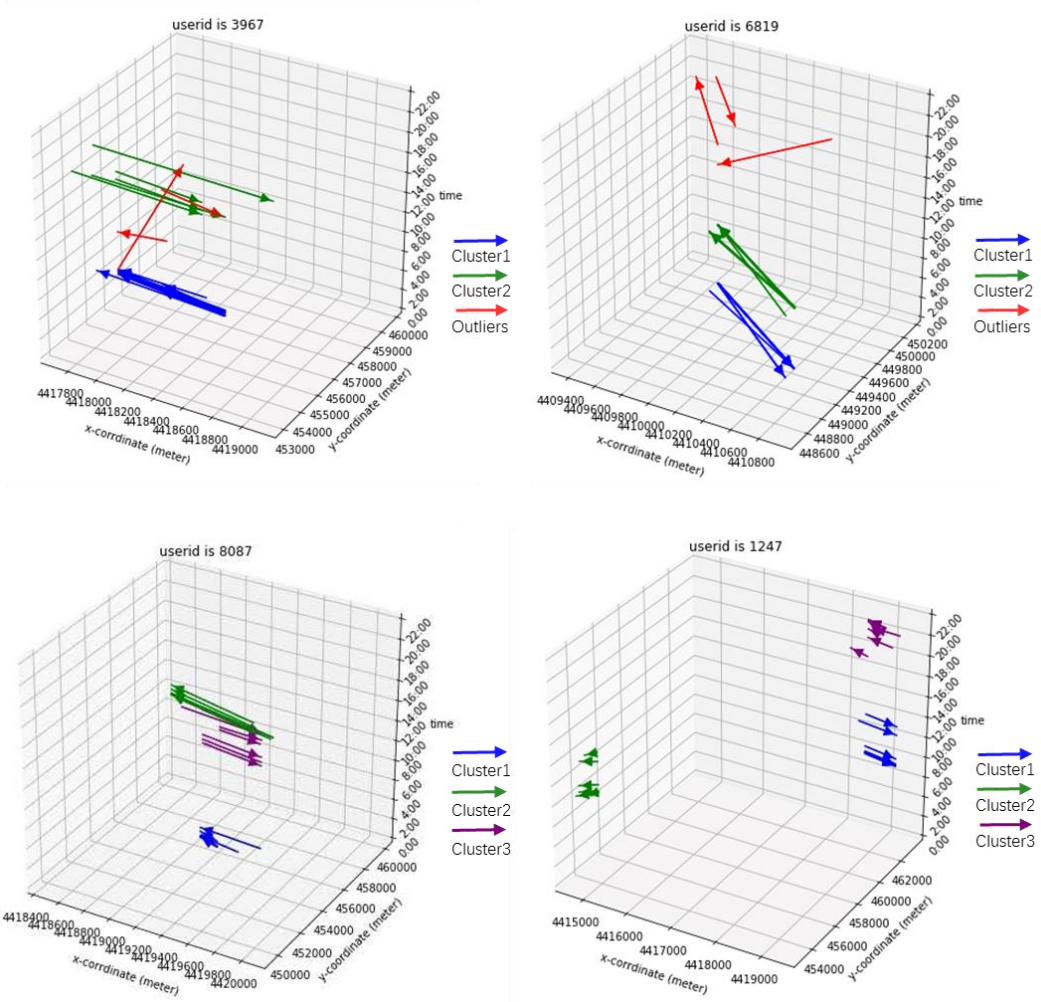

Fig. 4: Some of the individual users' clustering results.

Separating the individual users' regular demand and random demand from spatial-temporal trajectory clustering results, and then, identify the total regular demand that will recur of an area on the day to be predicted and predict total random demand based on MLP neural net model of the corresponding area. At last, calculate the sum demand of 10 areas respectively by adding identified total regular demand and total random demand together. Fig. 5 shows the final experimental results. As shown in the figure, the ten areas selected in the experiment are from Fengtai District, Beijing. Among the 10 areas, the largest area is 1.27 km^2 and the smallest one has only 0.18 km^2. Visualized results of predicted demand and real demand for the two days of May 19 and May 23 are displayed in each area. The maximum prediction error in ten areas is 102 and the minimum error is just 4.

Besides, in order to evaluate the feasibility and accuracy of proposed methodology, we also carried out the comparative experiment using only the MLP Neural Net Model to predict. The comparative experiment adopt the

traditional way that most demand prediction related literatures used, it is, using historical total demand to predict future demand directly without separating regular demand and random demand. The mean absolute deviation (MAD) of prediction prediction results of each model are shown in Table 1.

On the whole, the proposed methodology has higher prediction accuracy than using only MLP Neural Net Model, the average value of 10 areas' MAD is 42.33 and 54.56 respectively. For the results using proposed methodology, the maximum MAD among ten areas is 79.5 belongs to the largest area. The minimum MAD is just 8.5. Correspondingly, the results using MLP Neural Net Model only has the maximum MAD of 97 and minimum MAD is 21.5. Experimental results suggest that the proposed method is a feasible method that can improve the accuracy compare with traditional prediction model.

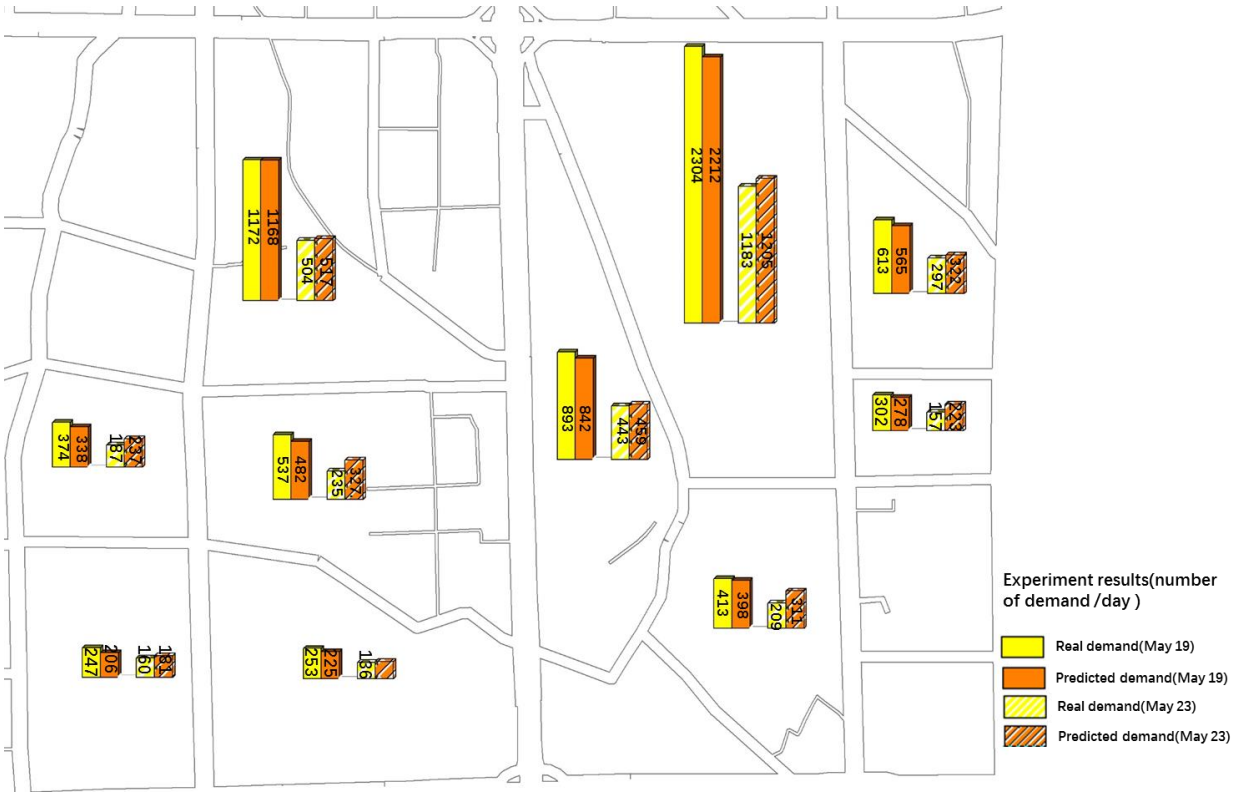

Fig. 5: Comparisons between predicted demand and real demand in ten areas.

Table 1: Comparisons of prediction results between proposed methodology and using MLP Neural Net Model only.

Area No.	Mean Absolute Deviation of two days' prediction results (MAD)	
	Proposed methodology	MLP Neural Net Model only
1	79.5	97
2	8.5	52.5
3	58.5	74.5
4	36.5	51.5
5	33	72
6	73.5	52.5
7	43	45.5
8	31	21.5
9	17.5	24
10	45	31

5. Conclusion

This paper contributes a novel method of predicting mid-long term FBBS travel demand, aiming to assist bike-sharing companies in determining the number of deployment and avoid the waste of resources and environmental hazards caused by blind expansion of shared bicycles. Specifically, a spatial-temporal distance function is determined and a spatial-temporal trajectory clustering based on Two Step algorithm was proposed for identifying the individual users' regular travel demand and random demand. Then, a MLP Neural Net Model was used for predicting total random demand based on the clustering results. The final predicted total demand equal to the identified regular travel demand plus the predicted random demand. At last, a numerical study using dataset from Beijing Mobike system is used for feasibility verification of the proposed methodology. By comparing the results with a comparative experiment using traditional historical total demand based prediction model, the proposed methodology in this paper has higher prediction accuracy.

However, as the very limited experimental data, we can only construct and validate a model of the prediction target for one day. Future work will collect larger and longer period FBBS dataset and build model with more variables for demand prediction of month, quarter and year.

Acknowledgements

This research was partly supported by the National Natural Science Foundation of China (General Program 51578199), which is gratefully acknowledged.

References

[1] X. Li, Y. Zhang, L. Sun and Q. Liu, "Free-Floating Bike Sharing in Jiangsu: Users' Behaviors and Influencing Factors," *Energies*, vol. 11, no. 7, pp. 1664, 2018.

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

[3] J. Corcoran, T. Li, D. Rohde, E. Charles-Edwards and D. Mateo-Babiabo, "Spatio-temporal patterns of a Public Bicycle Sharing Program: the effect of weather and calendar events," *Journal of Transport Geography*, vol. 41, pp. 292-305, 2014.

[4] C. Xu, J. Ji, P. Liu and L. Peng, "Forecasting the Travel Demand of the Station-Free Sharing Bike Using a Deep Learning Approach," in *Transportation Research Board 97th Annual Meeting*, Washington DC, United States, 2018.

[5] A. Pal, Y. Zhang and C. Kwon (2018, Jan 31). "Analysis of Free-Floating Bike Sharing and Insights on System Operations, Final Report," [Online]. Available: https://cpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/6/132/files/2017/03/ USF_YR1_ZHANG_ANALYSIS-OF-FREE-FLOATING1-1c5i2im.pdf

[6] Y. Shen, X. Zhang and J. Zhao, "Understanding the usage of dockless bike sharing in Singapore," *International Journal of Sustainable Transportation*, pp. 1-15, 2018.

[7] M. Bordagaray, L. dell'Olio, A. Fonzone and Á. Ibeas, "Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques," *Transportation research part C: emerging technologies*, vol. 71, pp. 231-248, 2016.

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

[9] X. Wu, F. Zhan, K. Zhang and Q. Deng, "Application of a two-step cluster analysis and the Apriori algorithm to classify the deformation states of two typical colluvial landslides in the Three Gorges, China," *Environmental Earth Sciences* vol. 75, no. 2, pp. 146, 2016.

[10] (2019). Multilayer perceptron [Online]. Available: https://en.wikipedia.org/wiki/Multilayer_perceptron

[11] M.H. Esf, M. Afrand, S. Wongwises, A. Naderi, A. Asadi, S. Rostami and M. Akbari, "Applications of feedforward multilayer perceptron artificial neural networks and empirical correlation for prediction of thermal conductivity of Mg (OH) 2–EG using experimental data," *International Communications in Heat and Mass Transfer*, vol. 67, pp. 46-50, 2015.