Research on Data Analysis and Dispatching Strategies of Shared Bicycles based on artificial intelligence technology

Yuxiang Zheng *

Guangzhou Foreign Language School, Guangzhou, China, 511455

* Corresponding Author Email: 18820088830@139.com

Abstract. In the context of the increasing popularity of shared bicycles, optimizing their dispatch during peak hours has become a crucial issue in urban traffic management. This study addresses the optimization of shared bicycle dispatch during peak hours, using data from Beijing. This paper began by cleaning and decoding a week's dataset, then defined block activity values for different time periods to identify popular blocks during rush hours. Due to the small size and large number of blocks, this paper applied the K-means algorithm to group these into 34 larger blocks, then calculated activity during peak periods to identify popular large blocks. This paper selected six key attributes for feature extraction: user ID, starting point coordinates, departure time, day of the week, and bike type. This paper used KNN regression and a fully connected neural network to predict the destination's coordinates. After performing K-fold cross-validation, the KNN model showed better performance, with an average validation loss of 0.0293 compared to the neural network's 0.0308. To address bike distribution issues, this paper proposed a manual adjustment strategy to balance bike availability across blocks. Using the 30th popular large block during the morning rush hour, we constructed a periodic adjustment method with 15-minute intervals. This paper simplified the problem into a constrained open traveling salesman problem and used enumeration and greedy algorithms to find the optimal adjustment strategy. This approach aims to stabilize bike distribution across blocks through spatial transfers. In conclusion, this study offers a framework for optimizing bike dispatch during peak hours and provides a scalable solution for managing bike distribution in larger urban areas during both rush hours and non-working days.

Keywords: Geohash Encoding and Decoding, K-means Clustering Algorithm, KNN Regression Model, Constrained and Open Traveling Salesman Problem.

1. Introduction

With the rapid development of urban transportation systems, shared bicycles have emerged as an essential mode of travel in cities worldwide. Their increasing popularity, especially during peak hours, has led to a need for effective dispatch strategies to optimize bike distribution and meet user demand. Efficient dispatching not only improves user experience but also enhances the overall operation of urban transport networks, reducing congestion and increasing the availability of bikes where needed most.

The optimization of shared bicycle dispatch has garnered significant attention in recent years, with several studies focusing on demand prediction, bike allocation, and the application of machine learning algorithms to enhance operational efficiency. One of the earliest approaches to this problem involved demand forecasting, with researchers utilizing historical usage data to predict bike demand in real-time. Xu et al. employed machine learning models to forecast bike demand at different locations based on temporal and spatial factors, such as time of day, weather conditions, and bike type [1]. Bao et al. explored clustering algorithms, including K-means and DBSCAN, to group geographically close stations with similar demand patterns, thus optimizing bike redistribution. These studies highlight the importance of predicting demand accurately to ensure the availability of bikes during peak hours. Further research focused on optimization techniques to balance bike supply and demand [2]. Hu et al. introduced a genetic algorithm-based optimization model to minimize the imbalance of bikes between stations [3]. Li et al. used simulated annealing to address the dynamic nature of bike-sharing systems [4]. However, many of these approaches have been limited to static optimization, where bike demand is considered as a fixed variable, neglecting the variability during

different times of the day or seasons. Such methods often fail to account for sudden fluctuations in demand that are characteristic of urban transportation systems during rush hours.

More recent studies have turned to dynamic models that adjust bike dispatch in real-time, considering the ongoing bike usage patterns and the potential for spatial transfers between stations. For example, some scholars introduced real-time optimization frameworks, leveraging reinforcement learning to learn optimal bike dispatch strategies based on current demand and supply [5-6]. This method allows for more adaptive and responsive solutions, yet it requires substantial computational resources and accurate data inputs to be effective. In conclusion, while much progress has been made in shared bicycle dispatch optimization, challenges remain in addressing dynamic, large-scale systems that consider both short-term fluctuations and long-term patterns. The combination of demand prediction, clustering algorithms, and real-time optimization strategies offers a promising direction for solving these problems. However, there is a gap in integrating these approaches in a unified framework that can manage bike distribution across large urban areas, particularly during peak hours. This study aims to fill that gap by proposing a comprehensive approach that includes data preprocessing, demand prediction, clustering, and a manual adjustment strategy based on spatial transfers, providing a scalable solution for bike dispatch optimization.

This study proposes an integrated approach to address these challenges. First, this paper performs data preprocessing, including Geohash decoding and cleaning, and define block activity values to identify popular areas during peak times. Next, this paper use K-means clustering to group small blocks into larger ones for more meaningful analysis. This paper then apply regression models, specifically KNN and a fully connected neural network, to predict bike destination coordinates. Finally, this paper develops a manual adjustment strategy based on the optimal bike transfer method, modeled as a constrained open traveling salesman problem, to maintain balanced bike distribution. This approach provides a scalable solution for optimizing bike dispatch during peak hours and can be applied to other urban environments.

2. Methods

2.1. Geohash encoding and decoding

The basic principle of GeoHash is to consider the Earth as a two-dimensional plane and recursively divide the plane into smaller sub-blocks, with each sub-block having the same code within a certain range of longitude and latitude [7]. Based on the GeoHash encoding principle, the encrypted codes of the blocks where the start and end points of the shared bike ride data are located are decoded. This involves converting Base32 encoding to decimal, decimal to binary sequence, separating odd and even bits to obtain the binary string of latitude, and using the binary search method to obtain the longitude and latitude intervals. Specifically, the geohash function can be used for reverse decoding to obtain the original longitude and latitude values of the points. At the same time, a polygonal grid coordinate can be obtained, representing the small block represented by this hash code. The processing of latitude and longitude information in the paper was carried out according to the above steps.

2.2. K-means clustering algorithm

K-means is a commonly used clustering algorithm [8], whose core idea is to divide the dataset into K clusters, so that the sum of distances between each data point and the centroid of its cluster is minimized. The core formula is as follows:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - u_i||_2^2 \quad (1)$$

The Silhouette Coefficient is a measure of cluster quality that assesses the tightness and separation of clusters. The contour coefficient is used to judge the rationality of clustering results by combining the similarity of samples within clusters and the separation degree of samples between clusters. If the contour coefficient is close to 1, it means that the sample is far away from other clusters in its own

cluster, the tightness within the cluster is high, and the separation between the clusters is good. If the contour coefficient is close to 0, it means that the sample is on the cluster boundary, the tightness within the cluster is not high, and the separation degree between the clusters is not high. If the contour coefficient is close to -1, it means that the sample is wrongly assigned to other clusters, and the tightness within the clusters is not high, and the separation degree between the clusters is poor. This paper divides peak areas by clustering different regions, with specific steps as described above.

2.3. KNN regression model

K-Nearest Neighbors (KNN) is a basic supervised learning algorithm for classification and regression problems [9-10]. When classifying or predicting new data points, KNN algorithm makes decision based on its similarity with the nearest neighbor data points in feature space. For regression problems, KNN algorithm will find the nearest k training data points (nearest neighbors) to the new data point when predicting the value, and then use the average (or weighted average) of these k nearest neighbors as the predicted value of the new data point, the core formula is as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (2)$$

Where, k is the number of neighbors of the predicted new sample. The choice of k will affect the complexity of the model and the prediction results.

3. Results

3.1. Data clean and identification of the popular residential blocks

By decoding the geohash, this paper obtained the latitude and longitude information represented by each code. Through observation, this paper found that there were a few outliers (with latitude and longitude information significantly different from the majority of the data) in the train data. By statistically analyzing the frequency of geohash codes appearing in the "geohashed start loc" and "geohashes end loc" columns of the train dataset, this paper delete the outlier's data.

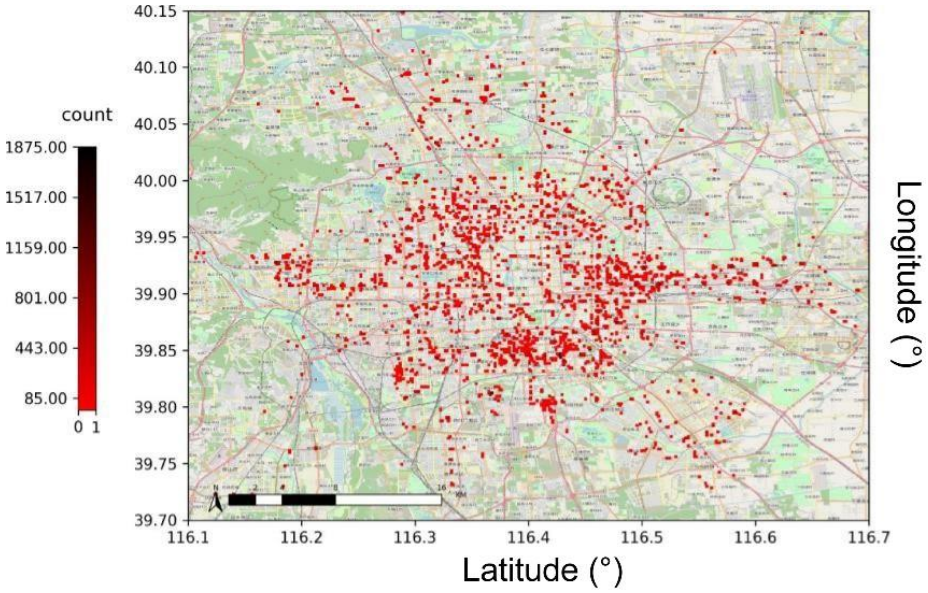

Figure 1. Morning peak activity heat map

To identify popular blocks, a new concept is defined for lock activity. The block activity of a block within a certain period of time is equal to the sum of the number of times it serves as the starting point and the number of times it serves as the destination during that period. For the convenience of subsequent discussions, this paper defines a small block as a rectangular grid obtained through hash encoding and decoding. To identify the popular grid blocks during the morning rush hour, this paper considered selecting all the data in the train dataset that falls within the morning rush hour period,

and then calculated the block activity for each grid block based on this data. Finally, the top ten grid blocks with the highest block activity were selected as the popular grid blocks during the morning rush hour. This paper also visually presented the block activity of each grid block on a world map in a manner similar to a heat map. The same approach was applied to determine the popular grid blocks during the evening rush hour, as shown in the Figure 1.

3.2. Hot large blocks recognition based on clustering algorithm

Given that the area represented by the above cell blocks is too small, resulting in excessive block division and too small amount of data in each block, which is not conducive to the actual management of operators, it is considered to combine K-means clustering analysis method to divide part of the cell blocks into the same large block, so as to facilitate the systematic management of subsequent operators with large blocks.

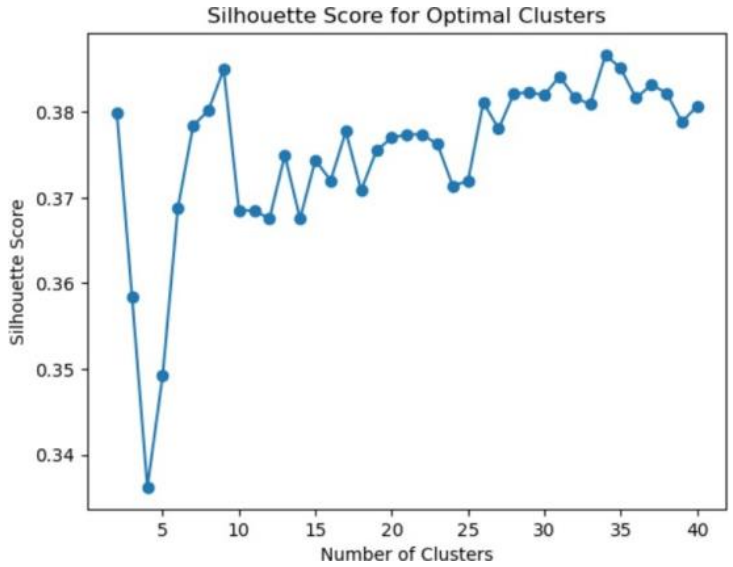

Figure 2. Contour coefficients corresponding to clustering numbers 2 to 40

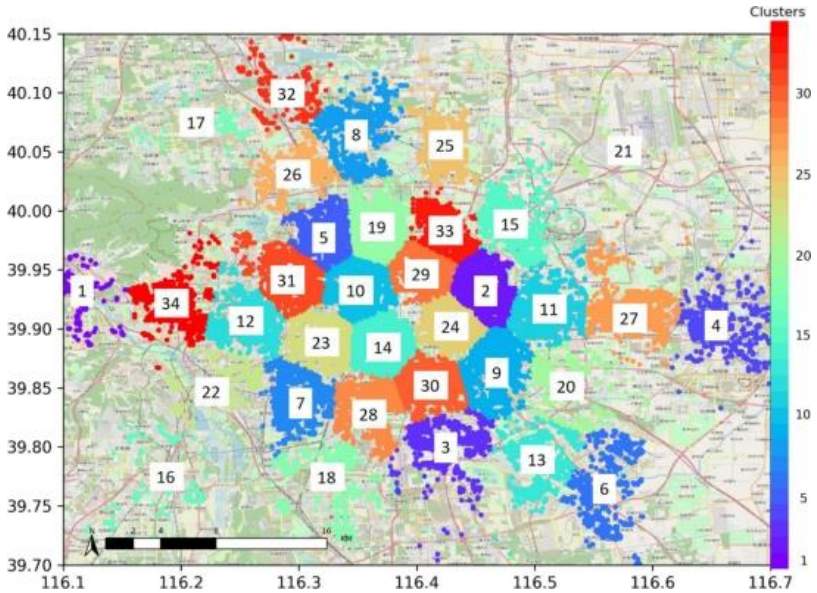

Figure 3. Large block partition diagram

Considering the management mode of shared bike operators and the relative stability of core block distribution, the clustering division is directly derived from all the data in the evening peak of the train dataset. Relatively speaking, clustering the evening peak into regions, or clustering the data of different weeks into regions, will lead to changes in regional division due to different data, which is

not conducive to operator management, nor is it conducive to the subsequent unified division standards to determine hot large blocks. This paper substituted the longitude and latitude of all the small blocks that occurred in the morning and evening peak of the train dataset into the calculation of contour coefficients, and calculated the contour coefficients corresponding to the clustering number from 2 to 40 respectively. The cluster results can be found at Figure 2, it can be seen that when the cluster number is 34, the clustering effect is best (with highest silhouette score).

The clustering parameters and the longitude and latitude of all the small blocks that occurred in the morning and evening peak of the train dataset were substituted into K-means clustering, and the clustering results were visualized to obtain a large block division diagram, as shown in Figure 3. In order to find popular large blocks during morning peak hours, we consider selecting all the data in the train dataset during morning peak hours, and calculate block activity for each large block based on this data: For all the cell blocks of the same large block (the grid corresponding to the hash code), this paper superimpose their block activity during the morning peak to become the block activity of the large block during the morning peak. Finally, according to the block activity, the top ten blocks are the hot blocks in the morning peak. The same can be said for the evening peak, as shown in Figure 4.

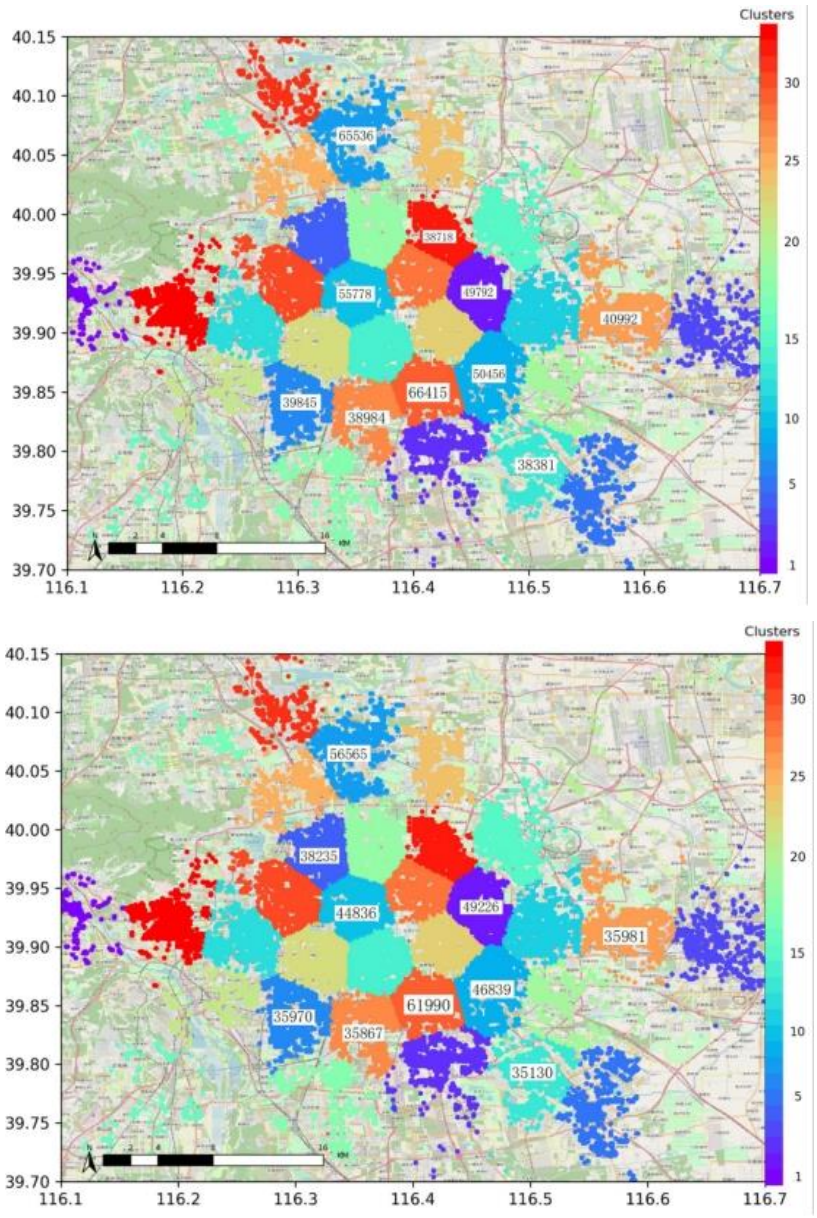

Figure 4. Top ten hot blocks in the morning and evening peak

3.3. Feature processing and prediction results

On the basis of the first part, the data screening is carried out first, and the data located in the popular large block in the morning peak period of the train data set is screened out as the training set. Extract feature attributes, including user number (userid), starting point longitude and latitude (lon-geohash-start, lat-geohash-start), departure time (starttime, weekday), and biketype (biketype), a total of six attributes. Use fit-transform to initially fit and transform the data, normalize the data and transform the data type.

Take the above six attributes as inputs and the latitude and longitude of the destination as outputs. The first layer is connected to the first hidden layer with 128 nodes. Then this paper has another hidden layer with 64 nodes and finally a connection between the second hidden layer and the output layer (with 2 nodes). Then design the network architecture and define how data flows through the network. Next, this paper set up an optimizer and loss criteria, set the learning rate to 0.001, then train the network, and finally complete the prediction on the test set. The prediction results are reversely normalized to the original coordinates, and the latitude and longitude values are converted to 7-bit geohash coding according to the coding rules of geohash, and then the hot block required by the topic can be obtained. The forecast results are shown in Table 1:

Table 1. Partial Prediction Results of KNN Regression Model

Longitude	Latitude	Weekday	Predicted	Predicted	Predicted	Geohash	Cluster	Weekday
116.53954	39.920883	5	116.51387	39.922523	wx4g57d	wx4g57d	11	116.53954
116.38161	39.849472	7	116.37	39.851498	wx4f8nd	wx4f8nd	28	116.38161
116.39397	39.859085	7	116.39764	39.862534	wx4fb8y	wx4fb8y	30	116.39397
116.33904	39.918137	4	116.33492	39.916122	wx4ep68		0	116.33904
116.39397	39.842606	6	116.39568	39.845707	wx4f8ts	wx4f8ts	30	116.39397
116.17699	39.914017	3	116.19854	39.917942	wx4eh4v	wx4eh4v	34	116.17699
116.464	39.876938	5	116.47183	39.88151	wx4ffkh		0	116.464
116.45851	39.864578	2	116.45789	39.867413	wx4ff1d	wx4ff1d	9	116.45851

Since there are more data in the training set (307,183), this paper takes k=3 and use 3-fold cross test to test the prediction model. Defines the mean square error loss function, which calculates the mean square error between the predicted output and the actual output, i.e. the average of the square of the difference between the predicted value and the actual value:

$$loss(x, y) = \frac{1}{n}\sum(x_i - y_i)^2 \tag{3}$$

Where x represents the predicted value and y represents the actual value.

As the results, both models have good effects, with mean square error of about 0.03. Compared with KNN, the effect is slightly better. Therefore, the results of KNN regression model are used as the prediction results of the hot block in the test data set where the cycling destination is located during the morning peak hours.

3.4. The optimal path planning

Before the modeling, the hypothesis of the model is explained as follows: through the path tracking of the same bike ID, this paper find that the existing bike scheduling mode has the scheduling characteristics of short distance and many times, as shown in the figure below. Therefore, this paper assume that the operator performs manual adjustment in the unit of the large block constructed in the first question, and the adjustment is only carried out among the cell blocks contained in the large block (the area represented by a hash code), and it is assumed that the manual adjustment is carried out once every 15min, and the completion time of the bicycle transportation process is 15min.

Because this paper needs to make adjustment decisions for hot blocks, in order to facilitate modeling and calculation, this paper take Block No. 30, which is the most popular block in the morning peak period obtained in the first question, as a typical example, and make adjustment

decisions in its morning peak period. For other popular blocks and other periods (such as evening peak), the model established in this paper has extensibility and can be used in the same way.

During data preprocessing, it is found that the data difference between working days and non-working days is large, while the difference between the two is small. Therefore, the two are considered separately in the calculation of the model, and the modeling methods of both are the same. This paper averaged 11 days of this paperekday data out of 15 days of data as a forecast of orders to guide weekday scheduling decisions.

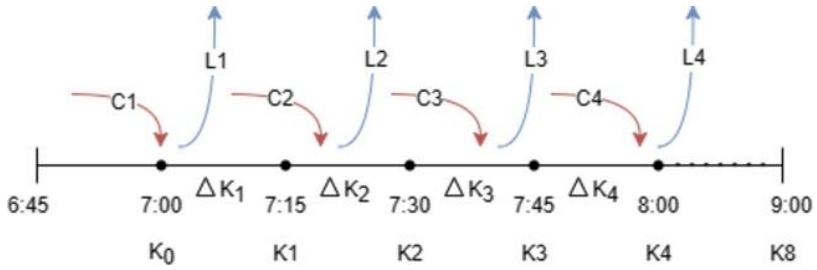

Figure 5. The change diagram of the number of bicycles in the KTH block

For any KTH cell block, there is an initial number of bicycles K_0 at the initial time node. At the other eight-time nodes in the morning peak, the operator calculates the changes in the number of bicycles in this block and gives the statistical result K_i as the data basis for scheduling decisions every 15min. According to the model hypothesis, since the average cycling time of bicycles in a large block is 15min and each time period is exactly 15min, the increase of bicycles in block K during this period is equal to the number of bicycles departing in the previous period and the destination is block K. The reduction of K-block bikes in this period is equal to the number of bikes departing from K-block during this period. Therefore, the change of bicycles in a period of time is equal to the increase of bicycles minus the decrease of bicycles, and finally, the number of bicycles at this time node is equal to the number of bicycles at the previous node plus the change of bicycles at this period of time, so the statistics of bicycles at another eight-time nodes in the morning peak can be derived. The number of bicycles in the KTH block changes as shown in the Figure 5 and equation (4) and (5):

$$C_i + L_i = \Delta K_i, i = 1,2,3,\ldots,8 \tag{4}$$

$$K_i = K_{i-1} + \Delta K_i, i = 1,2,3,\ldots,8 \tag{5}$$

It can be seen from the hypothesis that manual adjustment is carried out once every 15min, and the completion time of the bicycle transportation process is 15min. In other words, in the nine-time nodes and eight time periods constructed by us, each adjustment decision is carried out on the time node, and the adjustment process is exactly the next time period. That is, the last adjustment task is completed when node statistics are collected at the next time. If the T_{ki} of the number of bicycles changed by artificial adjustment is added in any i period, the number of bicycles artificially adjusted in the previous period should be added to the statistics of the number of bicycles in any time node (if it is negative, it is taken away), as shown in the Figure 6.

$$T_{Ki} + \Delta K_i = 0, i = 1,2,3,\ldots,8 \tag{6}$$

$$(K_i)' = (K_{i-1})' + \Delta K_i, i = 1,2,3,\ldots,8 \tag{7}$$

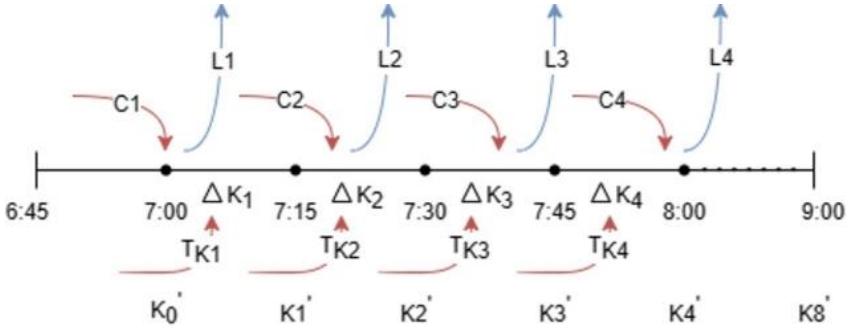

Figure 6. Diagram of the variation of the number of bicycles in the KTH cell block under manual regulation

According to the classification of large blocks obtained in the first question, this paper first selects the data related to the block in block 30 from the train data set (as long as the origin or destination is the block in block 30), and then filter out the data of working days. Since the train data contains 15 days from 2017.05.10 to 2017.05.24, in order to consider the scheduling decision of general working days, a total of 11 days of data of all working days are selected, and the subsequent values will be divided by 11 to represent the general working days. Select the data that starts between 7:00 and 9:00, and divide the data into eight periods according to the rule of 15 minutes. Then the data of each period is classified according to the hash code of each cell block, and the variation of each cycle in each period of each cell block is calculated according to the above modeling rules.

Combined with the above calculation results, this paper finds that the average change of bicycles in each period of the cell block in state 2 is generally 13 (±10), while the average change of bicycles in the cell block in state 1 is generally -17 (±5). In view of the limited loading capacity of general transport vehicles, and the limited reserve of bicycles on transport vehicles, it is necessary to attach more constraints to this unclosed travel salesman problem according to the actual situation: three blocks of the same state cannot be passed continuously during the transport process. It should be noted that considering that the true length of the transportation route between two points is generally the Manhattan distance rather than the European distance, the Manhattan distance is used to represent the distance between two points in the solution process.

For the unclosed traveling salesman problem with constraints, this paper use enumeration method and greedy algorithm to solve. Based on the idea of exhaustion, the program tries all combinations of paths that satisfy the constraints, then calculates the total distance of each path, and finally finds the shortest path. Since the algorithm tries all possible paths, it is guaranteed to find the shortest path, but the computational complexity increases dramatically as the number of points increases.

As a heuristic algorithm, the more common is the nearest neighbor algorithm, which starts from every starting point in state 2, selects the unvisited point closest to the current position each time the constraint is satisfied, and continues until all points have been visited. Although greedy algorithms do not necessarily find optimal solutions, they generally have low computational complexity.

Finally, combined with the schematic diagram of the optimal transportation route, this paper can get the bicycle scheduling strategy at 7:15 hours of a normal working day as follows:

The transport vehicle initially carrying 13 bicycles starts from the wx4f9mu cell block, and 23 bicycles are stored at the beginning. Go to the wx4f9ms block and deposit 13 bikes. Go to the wx4f9mk block and deposit 17 bikes. Head to the wx4f9kn block and drop off 21 bikes. Head to the wx4f9ky block and drop off 21 bikes. Go to the wx4f9mw block and deposit 7 bikes. Go to the wx4f9wb block and drop off 16 bikes. Go to the wx4fc9k block and drop off 13 bikes. Go to the wx4f9jd cell block and deposit 10 bikes. Finally, go to the wx4fb9x block and drop off the 12 bikes while the transporter is empty.

4. Conclusion

This study provides a comprehensive and practical approach to optimizing shared bicycle dispatch during peak hours. In the first part, the identification of popular large blocks using K-means clustering offers a solution that effectively addresses the limitations of small Geohash blocks, making the analysis both comprehensive and applicable for urban traffic management, the clustering results of 34 categories show noteworthy hotspots, which are used to assist in subsequent planning tasks. In the second part, the comparison of two prediction models (KNN and fully connected neural network) using K-fold cross-validation provides a convincing evaluation of their performance, with the KNN model proving to be more reliable. This robust comparison enhances the credibility of the results. In the third part, the simplification of the manual scheduling model offers a more efficient decision-making process, reducing decision time while providing clear objectives for bike redistribution. The optimal path planning results provide feasible scheduling strategies.

However, the model does have its limitations. The scheduling policy in the third part does not account for the status of all blocks, leading to suboptimal solutions for some situations. Despite this, the decision-making approach demonstrated in the third part can be extended to any large block, any time period, or non-working days, providing a scalable and adaptable solution for shared bicycle management. Future work could address these gaps by incorporating broader modeling frameworks and more comprehensive scheduling strategies.

References

[1] Xu C, Ji J, Liu P. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets [J]. Transportation research part C: emerging technologies, 2018, 95: 47 - 60.

[2] Bao L, Liu Z, Miao R, et al. Spatiotemporal clustering analysis of shared electric vehicles based on trajectory data for sustainable urban governance [J]. Journal of Cleaner Production, 2023, 412: 137373.

[3] Hu R, Zhang Z, Ma X, et al. Dynamic rebalancing optimization for bike-sharing system using priority-based MOEA/D algorithm [J]. IEEE Access, 2021, 9: 27067 - 27084.

[4] Li X, Wang X, Feng Z. Dynamic repositioning in bike-sharing systems with uncertain demand: An improved rolling horizon framework [J]. Omega, 2024, 126: 103047.

[5] Zhou Y, Li Q, Yue X, et al. A novel predict-then-optimize method for sustainable bike-sharing management: a data-driven study in China [J]. Annals of Operations Research, 2022: 1 - 33.

[6] Zheng Y, Hao Q, Wang J, et al. A Survey of Machine Learning for Urban Decision Making: Applications in Planning, Transportation, and Healthcare [J]. ACM Computing Surveys, 2024, 57 (4): 1 - 41.

[7] Petrov P. Practical approach for modifying existing geocoding system from equal angular to equal area[J]. Economics and computer science, 2023, 9 (2): 43 - 65.

[8] Ikotun A M, Ezugwu A E, Abualigah L, et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data [J]. Information Sciences, 2023, 622: 178 - 210.

[9] Abu Alfeilat H A, Hassanat A B A, Lasassmeh O, et al. Effects of distance measure choice on k-nearest neighbor classifier performance: a review[J]. Big data, 2019, 7 (4): 221 - 248.

[10] Halder, Rajib Kumar, et al. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications [J]. Journal of Big Data, 2024, 11 (1): 113.