Characteristic Analysis and Prediction Modeling of Car-sharing User Rental Based on Fisher Ordered Clustering

Ru Zhi¹, Jun Zhang², Jun Bi*, Zhen Zhuang¹, Qiu-Yue Sai¹
¹ School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China
² Yunnan Travelsky Airport Technology Co. Ltd, Kunming, China
*Corresponding author’s e-mail: bilinghc@163.com

Abstract. This work explores the characteristics of the rental behavior of Carsharing users on the basis of actual operation data from a car-sharing company in Beijing, China. Considering the random fluctuation of carsharing rental, the original data are clustered by fisher ordered clustering algorithm, and one day is divided into five time periods. Then we construct the SARIMA model for all data and the combination model of five SARIMA in different time periods. Finally, the evaluation indicators are calculated to compare the two modeling effects. The result is that establishing SARIMA prediction model after the use of fisher ordered clustering algorithm to divide the time period is better and the prediction accuracy is higher.

1. Introduction
Car-sharing is a kind of transportation mode in which consumers obtain the right to use the car in a short period of time by borrowing the car from the car-sharing operators to meet their own needs of motorized travel. The commercial operation of car-sharing has a history of nearly ten years in some developed countries, and it has just sprung up in China in recent years.

As a new type of transportation mode, the research results on the rental characteristics of car-sharing users are relatively few. Schmöller, S and Weikl, S constructed a linear regression model to analyze the factors affecting the demand for sharing cars, after clustering social demographic characteristics, predetermined time and spatial distribution. It is concluded that appointments are mainly concentrated in specific time periods and regions. The changes of weather conditions affect short-term demand forecasting, and the characteristics of social population affect long-term demand forecasting [1]. Kang, J, and Hwang, K constructed a multiple linear regression model to determine the factors affecting the use of site-based sharing cars, taking the traffic volume of car-sharing in Seoul as dependent variable, and the building environment, demographic and traffic variables as independent variables. It is concluded that there is a higher demand for car-sharing in areas with a higher proportion of building area or a larger number of sharing cars and fewer subway entrances, and that the proportion of young residents aged 20 to 30 who use car-sharing is higher [2]. Müller, J and Correia, GHD constructed a negative binomial statistical model to predict the number of reservations for free floating car-sharing, using the reservation data provided by DriveNow for Berlin and the type of each area. It is concluded that car-sharing is more easily promoted in regions with more affluent citizens, and the availability of central areas and parking lots affects the reservations of car-sharing [3]. Stasko, TH and Buck, AB studied the relationship between the parking demand of car-sharing and the time and purpose of travel by using online questionnaires and parking data from the University of Ithaca, New York. It is concluded that geographical areas and parking types affect parking demand [4].
Ying H.I and Wei, W concluded that the rent of users presents three peaks in the morning, mid and evening, based on Hangzhou Chefenxang order data, however there is no in-depth modeling and analysis [5].

The existing literature mainly focuses on the external factors affecting the number of rental cars, but seldom extracts the features of the rental car data itself. Due to the orderliness and periodicity of the rental car data itself, the prediction accuracy can be improved by selecting the appropriate data clustering algorithm to extract the data characteristics. Therefore, this paper uses fisher ordered clustering algorithm to divide the rental period of car-sharing users, and then proposes SARIMA model to forecast the number of rental cars in different time periods. By comparing the model constructed by the data before and after clustering, it is found that the prediction effect of clustering data is better, which will provide more precise guidance for the operation and scheduling of car-sharing operators.

2. Car-sharing data analysis

2.1. Dataset
The data are from a car-sharing company in Beijing, China, which was established in 2015 and provides one-way station-based car-sharing service. The user is allowed to return a rented car to any car-sharing station, which may not be the origin station. The data contain the car-sharing rental order data of the company from May 2017 to August 2017 with a total number of 49849.

2.2. Rental and Return Analysis of User
To analyze and compare the distribution of users’ rental and return of cars on the daily time axis clearly, we draw the daily average of renting and returning cars on the time axis of one day, as shown in Fig. 8, where a day is divided into 48 30-minute time periods. The car rental curve has three peaks, namely, 9:30 am, 12:00 noon and 5:30 pm. The growth of morning peak travel is slower than that of evening peak travel, because travel time point of users at morning peak is different but the travel time point of users at evening peak is concentrated. The peak time of returning curve is around 9:00 am. Using the number of renting and returning cars in each period at each station and the car number of parking spaces, the operators can reasonably dispatch the cars. For the users, it is possible to understand the usage demand of the other users in advance, and is beneficial to plan the travel time in advance.

3. Clustering analysis based on fisher ordered clustering algorithm

3.1. Introduction to the principle of fisher ordered clustering
Fisher ordered clustering algorithm, also known as optimal division algorithm, was proposed by Fisher in 1958. The principle of clustering is to minimize the difference between the ordered samples of each group and maximize the difference between each group. Different from general clustering analysis, it classifies ordered samples, which means that samples must be adjacent in order to be in the same class.
Step 1: define class diameter
Suppose the sample composition of class CK after clustering is \{Xi, Xi+1, …, Xj\}, D(i, j) is the diameter of the class, where \( \bar{X} \) is class mean vector of CK:

\[
D(i, j) = \sum_{t=i}^{j-1}(X_t - \bar{X})
\]  

(1)

Step 2: define loss function of classification
Assuming that n samples are classified into k categories, the ordinal number of the first sample of each category is the segmentation point, and the ordinals are I = s1, s2, s3, …, sk, then the loss function \( L(n, k) \) of this classification is the sum of diameters of all classes:

\[
L(n, k) = \sum_{t=1}^{k} D(s_t, s_{t+1} - 1)
\]  

(2)

When n and k are fixed, the smaller the value of loss function is, the more reasonable this classification is. When n and k are fixed, the minimum loss function value is \( L(n, k)* \). According to the meaning of the following two recursive formulas:

\[
L(n, 2) = \min_{i \in [1:n-1]} \{D(1, i - 1) + D(i, n)\}
\]  

(3)

\[
L(n, k) = \min_{i \in [1:n-1]} \{L(i - 1, k - 1) + D(i, n)\}
\]  

(4)

Step 3: seek the optimal solution
When sample number n and classification number k are fixed, the minimum loss function value \( L(n, k)* \) can be obtained according to formula (3) and (4), and the classification corresponding to \( L(n, k)* \) is the optimal solution.

3.2. Empirical Study
Taking 30 minutes as the time interval, a day is divided into 48 time periods, and the number of rental cars per day is also divided into 48 orderly sample data. Considering that the number of rental cars in the period from 23pm to 24pm in a day is less and is more similar to those in the early morning, so we sum up the data of the last hour of the previous day and the data before 23pm on that day as one-day statistics, and all data are processed by analogy. Firstly, data points are normalized. Because the number of rental cars shows periodicity and daily similarity, this paper takes the normalized data of the average number of rental cars in the same period as a sample, and constructs the input vector \( X_r = \{X_1, X_2, …, Y_{r2}\} \), set the optimal number of classifications K in the range of 2-10.

3.2.1. Calculating class diameter

| i | j | 1   | 2   | 3   | ... | 46  | 47  | 48  |
|---|---|----|----|----|----|----|----|----|
| 2 |   | 0.0015 | 0 | 0 | ... | 0 | 0 | 0 |
| 3 |   | 0.0043 | 0.0007 | 0 | ... | 0 | 0 | 0 |
| 47 |   | 4.1862 | 4.2181 | ... | 0.0005 | 0 | 0 | 0 |
| 48 |   | 4.2327 | 4.0473 | 4.1465 | ... | 0.0063 | 0.0030 | 0 |
3.2.2. Calculating the minimum classification loss function

Table 2. Minimum classification loss function $L[p(n, K)]$

| $k$ | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 3   | 0.001 (2) | 0 (0) | 0(0) | 0(0) | 0(0) | 0(0) | 0(0) | 0(0) | 0(0) |
| 4   | 0.003 (3)  | 0.001 (4) | 0(0) | 0(0) | 0(0) | 0(0) | 0(0) | 0(0) | 0(0) |
| ... | 3.953 (43)  | 0.719 (44) | 0.309 (43) | 0.225(43) | 0.187(44) | 0.154(44) | 0.124(44) | 0.097(44) | 0.078(45) |
| 48  | 3.964 (43)  | 0.733 (44) | 0.332(43) | 0.248(43) | 0.202(44) | 0.168(44) | 0.139(44) | 0.112(44) | 0.085(46) |

3.2.3. Determining the optimal classification number $k$. Make a trend chart of $L[p(n, K)]$ changing with $K$, as shown in Figure 2. From the graph, it can be seen that the curve begins to turn and tends to be flat at $K=5$ and $6$. In order to ensure that the loss function is as small as possible and the number of intervals is as small as possible, $K$ is chosen as $5$.

![Fig 2. The minimum loss function value under different classification number](image)

![Fig 3. The results are based on the state interval of the fisher ordered clustering](image)

Therefore, this paper divides a day's car rental data into five sections, the results of which are shown in Fig 3 and Table 3.

Table 3. The result of the partition of the state interval

| Section | Interval period   | Section name         |
|---------|-------------------|----------------------|
| Section 1 | 23:00-07:30      | Trough of night    |
| Section 2 | 07:30-9:00       | Morning peak       |
| Section 3 | 9:00-17:30       | Daytime peak period |
| Section 4 | 17:30-20:30      | Evening peak       |
| Section 5 | 20:30-23:00      | Decline of night   |

4. Prediction and Contrast Analysis Based on SARIMA

4.1. Introduction to the SARIMA model

SARIMA model is a special autoregressive integrated moving average model (ARIMA). If there is only a strong growth trend and no obvious seasonality in the time series, the original sequence can be processed by the low-order difference process, and the ARIMA model can be constructed to predict accurately the stationary difference sequence. But for the number of rental cars, which has both trend growth components and obvious seasonal fluctuation cycle series, needs to add another cycle difference in ARIMA model to construct a SARIMA model to forecast[6].

The expression of ARIMA model is ARIMA $(p, d, q)$, where $p$ represents the lag order of autoregressive operator, $q$ represents the lag order of moving average operator, and $d$ represents the number of differences. The general expression of ARIMA $(p, d, q)$ model:

$$\Phi(L)\Delta^d y_t = \Theta(L) u_t$$  \hspace{1cm} (5)
\( \Phi(L) \) said smooth handover, \( \Theta \langle L \rangle \) said reversible moving average operator, \( d \) said difference frequency, \( u_t \) said random disturbance, and \( u_t \sim \text{iidN}(0, \sigma^2) \). However, in some ARIMA models, there are obvious seasonal variations. This seasonality is caused by climate change, business practices and expectations, or other factors, and is periodic over a certain period of time. Therefore, the sequence with periodic changes is called seasonal sequence. Therefore, seasonal ARIMA model was needed to describe and represent this type of sequence with SARIMA, which was also called multiplication seasonal model for earlier literatures. The general expression of seasonal time series model:

\[
\Phi_p(L)A_p(L^s)(\Delta^d\Delta^s y_t) = \Theta_q \langle L \rangle B_q v_t
\]

(6)

P, Q, P, Q, respectively season with the seasonal autoregressive and moving average operator is the maximum number of delayed order, \( A_p(L^s) \) and \( B_q(L^s) \) season respectively P order autoregressive moving average operator, operator and Q order season \( \Phi_p(L) \) and \( \Theta_q \langle L \rangle \), respectively, said the ar (P) operator and the Q order moving average operator, \( d \), D said the season and the number of seasonal difference, \( v_t \) said random disturbance, and \( u_t \sim \text{iidN}(0, \sigma^2) \). Formula (2) is called \( (p, d, q) \times (P, D, Q) \)s order seasonal ARIMA model or Multiple Seasonal ARIMA.

4.2. Empirical Study

If we can know the change of the number of rental cars in different time periods of the next day in advance, it will provide data support for the operation and scheduling of the car-sharing operators. This paper chooses the data from May 2, 2017 to August 29, 2017 as training samples, and takes August 30, 2017 and August 31, 2017 as test samples to establish SARIMA prediction model to analyze the results.

4.2.1. Primitive sequence analysis and sequence stabilization

Draw the time series chart of the number of rental cars with 30 minutes interval in one day from May 1, 2017 to August 29, 2017, autocorrelation ACF chart and partial correlation PACF chart, as shown in Fig. 4 and Fig. 5. We can see from the Fig. 4 that the sequence has no obvious trend, but has a seasonal cycle of one day. It can also be seen from Fig. 5 that the autocorrelation decays slowly and peaks appear at the integer multiple of 48. After a long delay order, the autocorrelation can not converge near zero, which is the performance of non-stationarity. Table 4 is the result of ADF test on the original sequence, where \( p > 0.05 \), which also shows that the sequence is not stable and needs differential processing.

| Test critical values: | t-Statistic   | Prob.*     |
|----------------------|--------------|------------|
| Augmented Dickey-Fuller test statistic | -1.082023 | 0.3325 |
| 1% level             | -3.431465    |            |
| 5% level             | -2.862033    |            |
| 10% level            | -2.567032    |            |
The original sequence has seasonality, in which the period $S$ is 48. After experimentation, it is found that the difference sequence with one seasonal difference and one order difference is stable. The difference sequence is tested by ADF, that is, the sequence is stationary.

4.2.2. White noise test before modeling. According to the white noise test method, LB statistics are constructed and P values are calculated, as shown in Table 5. When the delay is 6, 12 and 18 orders in the table, $p < 0.05$, it is considered that the difference sequence is not a white noise sequence and needs to be modeled and studied.

| Delay order | LB statistic | Prob.* |
|-------------|--------------|--------|
| 6           | 1169.372     | 0.000  |
| 12          | 1179.014     | 0.000  |
| 18          | 1183.506     | 0.000  |

4.2.3. Model Recognition. Because of one seasonal difference and one first order difference of the original sequence, that is $D = 1$, $d = 1$, $S = 48$ in the model. Considering that most of the sequences can be fitted very well by using low-order models, $0 \leq Q \leq 1$, $1 \leq P \leq 2$ can be set initially. We can be seen from the autocorrelation ACF and the partial correlation PACF within the order of 40 after the differential sequence delay that both autocorrelation coefficient and partial correlation coefficient are trailing, in which the delay of autocorrelation coefficient tends to 0 after 7 orders and that of partial correlation coefficient tends to 0 after 27 orders, considering at the same time that most of the sequences can be fitted very well by using low-order models, so $1 \leq q \leq 2$, $1 \leq p \leq 7$ is set initially. After many times of modeling, it is found that the value of AIC and SC of model SARIMA$(2,1,2)(2,1,0)_{48}$ are the smallest and the fitting accuracy is the highest.

4.2.4. Parameter estimation and model diagnosis. The least squares method is used to estimate the parameters, and the significance of the parameters is tested. The $P$ value of the parameter statistics $I$ is less than 0.05. It is considered that the parameters of the model $SARIMA(2,1,2)(2,1,0)_{48}$ pass the test. The prediction equation is as follows:

$$\hat{y}_{48t} = (1 - 0.998763B^2)\epsilon_t (1 + 1.720941B + 0.959349B^2)(1 + 1.930816B^{48} + 0.956793B^{96})$$

The significance test of the model is to test the white noise of the model residual sequence. It is observed that the P value of LB statistic is greater than 0.05 when the delay order is 6, 12 and 18. It can be judged that the residual of the model $\{\epsilon_t\}$ belongs to white noise, that is to say, the law of the original data has been fully excavated.

| Delay order | LB statistic | Prob.* |
|-------------|--------------|--------|
| 6           | 4.120        | 0.661  |
| 12          | 7.171        | 0.846  |
| 18          | 17.330       | 0.501  |

4.2.5. Experimental results. After the above experiments, the forecasting model of the number of rental cars is obtained. The accuracy of the model is validated by the data of August 30, 2017, and the forecast comparison chart and error graph are obtained, as shown in Fig.6 and Fig.8, respectively.
From the above results, it can be seen that the fluctuation curve of the predicted value basically fits the fluctuation of the real value. In the case of more than 80%, the absolute error of the prediction of the number of rental cars in different time periods is within $\pm 3$, and the analysis of the remaining three error indicators is shown in Table 7.

### 4.3. Contrastive Analysis

In order to validate the effectiveness of Fisher ordered clustering, this study constructed seasonal time series models for five different time periods of data after clustering and predicted the rental cars data in the corresponding interval. The prediction period is August 30, 2017 and August 31, 2017. The comparison and prediction results of the models are given in Table 6.

| Model      | Emean | RMSE | RMSRE |
|------------|-------|------|-------|
| Date       |       |      |       |
| No clustering | 2017.8.30 | 2017.8.31 | 2017.8.30 | 2017.8.31 |
| Clustering | 2.8541 | 2.6250 | 1.6885 | 1.6373 |

| Clustering | 1.4759 | 1.5456 | 1.1457 | 1.2897 | 0.3629 | 0.3776 |

Through comparative analysis, it is found that using fisher ordered clustering to divide the data into different time periods to build models, which will get better prediction results. The application of this method to the prediction of rental and return of cars at various stations will provide strong data support for the operation and scheduling of car-sharing operators.

### 5. Conclusion

Analyzing and predicting the usage characteristics of car-sharing users is of great practical significance for operators. In order to improve the prediction effect of the number of rental cars, it is particularly important to process and analyze the original data. In this study, fisher ordered clustering algorithm is used to classify the original data with periodic differences, which reduces the volatility of data and the complexity of prediction. Considering the obvious periodicity of the original sequence, we choose SARIMA to build the model. By comparing the EMean, RMSE and RMSRE, it is found that fisher ordered clustering algorithm can effectively divide the time interval of data, and the prediction accuracy of clustered data can be effectively improved.

### Acknowledgment

The data used in this paper were provided by a Beijing travel-sharing company and supported by the National Natural Science Foundation of China (71961137008).

### References

[1] Schmöller, S., Weikl, S., Müller, J., Bogenberger, K. (2015) Empirical analysis of free-floating Car-sharing usage: The Munich and Berlin case, Transportation Research Part C 5634-51.

[2] Kang, J., Hwang, K., Park, S. (2016) Finding factors that influence Car-sharing usage: case study in Seoul. Sustainability 8 (8), pp. 709.

[3] Müller, J., Correia, G., Bogenberger, K. (2017) An explanatory model approach for the spatial
distribution of free-floating Car-sharing bookings: A case-study of German Cities. Sustainability 9 (7), pp. 1290.

[4] Stasko, T. H., Buck, A. B., Gao, H. O. (2013) Car-sharing in a university setting: Impacts on vehicle ownership, parking demand, and mobility in Ithaca, NY. Transport Policy 30 (4), pp. 262-268.

[5] Hui, Y., Wang, W., Sun, QL. (2016) A study on the use behavior characteristics of Car-sharing users: A case study of Hangzhou Che Fen Xiang. Communication Shipping 3 (05), pp. 24-29.

[6] Tang, YY., Zhu, XL., Li, L. (2019) Railway monthly passenger volume prediction based on SARIMA model. Journal of Transportation Engineering and Information, 17(01):25-32.