ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20210507172841
Original scientific paper

A K-means Group Division and LSTM Based Method for Hotel Demand Forecasting

Tianyang WANG

Abstract: The accuracy of hotel demand forecasting is affected by factors such as the completeness of historical data and the maturity of models. Most of the existing methods are based on rich data, without considering that single hotels may only obtain sparse data. Therefore, a K-means group division and Long Short-Term Memory (LSTM) based method is proposed in this paper. Guest types are introduced into the forecasting to provide reference for hotel's further decision-making. Using an example of 1493 hotels in Europe, we divide hotel groups and forecast the flow of leisure and business guests. The experimental results show that, compared with the benchmark models, LSTM can improve the forecasting performance of hotel group; compared with single hotels, the forecasting of hotel groups can effectively avoid inaccuracy caused by sparse data. The results can provide necessary reference for hospitality to make decisions based on guest types.

Keywords: group division; hotel demand forecasting; K-means; LSTM; sparse data

1 INTRODUCTION

As a modern service industry management method, revenue management can help hotels to maximum revenue through theory and practice. In hospitality, demand modeling and forecasting are always closely related to hotel revenue management, and are often applied to hotel procurement decisions, inventory control, business operations management and planning [1]. Demand forecasting is an important step in the decision-making process for hotel planners and managers. Accurate forecasting is the key. Revenue management and other similar strategies take forecasting as the basic element of pricing [2]. Without accurate forecasts, pricing errors will have a negative impact on the hotel's financial performance. However, it is not a simple task to get accurate forecasting result. Many factors, such as incomplete historical data, immature prediction model, change of consumer habits, and special date will affect the accuracy. Many efforts have been made to improve the accuracy by measuring the demand for hotel accommodation from different angles through various variables. Some use variables related to financial performance to measure demand, such as using statistical residuals to predict revenue per room [3]; using autoregressive integrated moving average and intervention analysis technique to forecast hotel performance [4]. Some also use variables related to the demand scale, such as introducing the bank exchange rate index into the hotel standard demand equation to predict the number of rooms sold [5]; introducing business sentiment indicators to predict the actual number of hotel arrivals [6]; forecasting the number of overnight based on search engine data and LSTM model [7]; using fuzzy c-means clustering algorithm to forecasting occupancy rate [8]. No matter from which point of view, forecasting hotel demand is essentially based on the analysis of data. Hotel demand forecasting methods include time-series, advance booking and combined models, which mostly rely on the historical data, such as bookings and number of overnight. The density of data is one of the important influencing factors for hotel demand forecasting [9].

The data of a single hotel may be sparse. Such as hotels that have just entered the market cannot produce complete data to provide reference for future operation decisions due to their short opening time and insufficient competitiveness; hotels with low exposure, located in remote location or not in the major business area always keep a low level of historical data value due to the small number of guests, which makes it difficult to dig out valuable information. The sparsity of historical data not only has a great impact on the accuracy of hotel demand forecasting, but also increases the difficulty of forecasting and leads to the instability of results. Few researches have considered this. Literature [10] proposed a method of group division to solve the problem of single hotel data sparsity. Hotel Group division divides hotels according to some features to form one or more groups. The hotel data in each group are integrated to forecast, that is, to forecast the accommodation demand of each hotel group. Because the hotels in the same group have similar features, they can often provide reference for each other. In addition, it can also help enterprises that want to build hotel projects to better assess the risk level.

With the improvement of living standards, people have higher requirements on the quality of hotel service. Different guest groups have different requirements. For example, leisure guests have high requirements for food and accommodation, business guests have high requirements for processing speed of laundry and other affairs, guests with mobility difficulties have high requirements for convenience of access, and the elderly and the sick have high requirements for disease emergencies. In order to attract more guests and increase revenue, the hotel should take different measures to cope with different guests. At present, the existing demand forecasting methods mostly analyze all guests as one type, cannot locate the accommodation requirements of different types. Therefore, we introduce guests types into the demand forecasting, focusing on the analysis of leisure and business guests, so as to play a guiding role in further decision-making for hotel planners and managers.

Based on the above analysis, we construct the K-means model to group 1493 luxury hotels in Europe, and construct the LSTM model with good predictive ability for complex nonlinear time series to forecast the demand of hotels. We forecast the monthly flow of leisure and business guests in the hotel group to provide reference for revenue management, business procurement, and operation strategy.

Compared with the existing work, the contribution of this paper can be summarized as follows:
(1) K-means clustering algorithm is used to group hotels to avoid inaccurate forecasting caused by sparse historical data.
(2) A hotel group guest flow forecasting model based on LSTM is constructed for the first time.
(3) Introduce hotel guest types and forecast the monthly flow of leisure and business guests respectively.
(4) The ANN and SVR are built as the benchmark models, and compared with the LSTM adopted in this paper to evaluate the accuracy of forecasting the monthly flow of single hotels and hotel groups. The evaluation metrics include mean absolute error (MAE), root mean square error (RMSE) and the mean absolute percentage error (MAPE).

The remaining sections of this paper are organized as follows. Section 2 provides an overview of the existing hotel demand forecasting methods and discusses their advantages and disadvantages. Section 3 analyzes the review data of 1493 luxury hotels in Europe used in this paper. Section 4 introduces the proposed hotel group division and demand forecasting method in detail. Section 5 describes the experiments and analyzes the results. Section 6 summarizes the work of this paper.

2 LITERATURE REVIEW

In the hospitality, accurate forecasting is the key to revenue management and other related strategies. Time series, advance booking and combined model are three methods commonly used in hotel demand forecasting [8, 11, 12, 13]. This paper is particularly interested in time series forecasting models, and will focus on the researches related to this method.

Time series model is widely used. It looks for time patterns in a single historical data series, such as trends, cycles, and seasonal fluctuations, and then models the patterns mathematically [12]. Simple time series models such as moving average (MA), exponential smoothing (ES) and regression, and family models of autoregressive moving average (ARIMA) are used to forecast the final demand based on historical days [11]. Andrewet al. [14] constructed Box-Jenkins and ES models to forecast the actual monthly occupancy rates of hotels in major central cities, and obtained accurate results, proving the important role of time series models in actual hotel operations and other applications (such as yield management); the MA model was used to forecast the number of arrivals in Choice Hotels and Marriott hotels, and was proved to be highly robust [2]; in order to improve the accuracy of the daily occupancy rate of a single hotel, ARIMA and the model combining ensemble empirical mode decomposition (EEMD) and ARIMA were built, the experimental results show that EEMD-ARIMA has improved the accuracy of forecasting, especially short-term [15]. Although the traditional time series forecasting models have a wide range of applications, it can easily lead to inaccurate forecasts when the time series is non-stationary. In addition, the construction of the model is complex and requires high statistical expertise. In recent years, machine learning and deep learning have been proven to have good demand forecasting capabilities and can effectively capture nonlinear and complex features in time series [16-19], but they have not been widely used in the hospitality. Given the reservation and occupancy history records, ridge regression, kernel ridge regression, multilayer perceptron, and radial basis function networks are constructed to forecast the daily occupancy rate, and the good forecasting performance is obtained [20]; Aliyevetc. [8] established the hotel occupancy forecasting system model based on fuzzy C-means clustering algorithm; Tsang et al. [21] proposed using gaussian process to forecast daily occupancy rate, they also constructed linear regression, ARIMA, support vector machine, random forest and other recently commonly used machine learning methods in the experimental stage. Results showed that the performance of machine learning is generally better than traditional time series models. Like the transportation data with diversity and typical characteristics of big data [22], the data in hospitality is growing rapidly and comes from many events. Due to the weak ability to deal with the increasing data, machine learning is limited in hotel demand forecasting. Zhang et al. [7] used the Internet search index based on LSTM deep learning framework based on LSTM to forecast overnight guest flow in Hainan Province from August 2008 to May 2019, and compared it with DBN, BPNN and C-LSTM, proving the advantages of LSTM in complex time series forecasting; Literature [23] constructed the LSTM model to forecast the actual monthly arrivals of a resort hotel in Portugal, and compared it with artificial neural networks (ANN) and support vector regression (SVR) models. The result showed that LSTM has better performance in capturing the nonlinear complex features of time series data. Affected by factors such as weather and economy, the guests flow often shows instability. The excellent performance of LSTM on unstable time series with fixed components has been proven in researches. This paper will further explore the performance of LSTM in hotel demand forecasting.

Advance booking model considers the arrival number of booking requests within the booking scope of a specific stay for one night. The idea is to estimate the incoming booking increment, and then aggregate these increments into the early implementation to obtain the final demand forecasting result [11]. This type of model is considered to be the most accurate method for forecasting final demand in short term. Advance booking model can be divided into an additive and multiplicative model. Additive model assumes that the number of bookings on a certain day before arrival is independent of the number of rooms ultimately sold, while the multiplicative model assumes that the number of future bookings depends on the current number of bookings. Athanasius et al. [24] constructed 8 variants of advance booking models to predict the number of arrivals in a given range of hotel, the result showed that multiplicative variation is better than additive variation. Since advance booking model only uses the existing booking data on a certain day and ignores the past data, Tse and Poon [25] described the advance booking curve as a quadratic function and applied it to the forecast of Hong Kong ICON hotel; Lee et al. [11] established three Poisson models to capture key features of booking arrivals, independent of the number of rooms ultimately sold, time-varying arrival rate, and based on the daily booking data of 69 major hotel chains in the United States, the Poisson

hybrid model was proved superior to the standard and linear regression model.

Combined model is usually based on the weighted average of the forecasting obtained from different methods and information sources [13], which is widely accepted in practice to improve the accuracy. Literature [26] summarized a variety of combined models. These models prove that combination of forecasting generated by different models can achieve higher accuracy. The work of Rajopadhye et al. [27] is a good example; they combined a time series model and an advance booking model, using Holt-Winters to achieve long-term forecasting of hotel rooms and a multiplicative advance booking model to achieve short-term forecasting.

Hospitality relies heavily on existing data such as sales or bookings to make accurate forecasting for final demand, especially when using models based on time series or advance booking curves. If the data is sparse, it will become very difficult to capture the rules. Generally speaking, hotel demand forecasting is the analysis process of the data held by the hotel. Through analysis, it can provide hotel operators and managers with decision-making reference and point out the future development direction. Each hotel has the responsibility of collecting and analyzing available data, and making its own forecasts. Although the forecasting of a single hotel will help to implement operational strategies, it also puts forward higher requirements on the available data, and sparse data will lead to a decrease in the accuracy of the prediction. In the above researches, almost all models are established based on the complete data. However, in reality, hotels that have entered the market soon or have low exposure rates only have sparse data. This paper adopts a hotel group division method proposed in [10] to solve the problem of inaccurate or difficult forecasting caused by sparse data of a single hotel, and improve the demand forecasting method on the basis of this research. Our improvements include two aspects: selecting the best K value (the key parameter of K-means) by elbow method, and using LSTM algorithm which is friendly to non-stationary time series for forecasting. In addition, types of hotel guests are diverse, and the analysis of all guests as a single type cannot provide a reference for the hotel's further decision-making, most existing researches have not taken it into consideration.

The weakness of the above researches can be summarized as follows:

(1) The potential sparsity of data for single hotels is not taken into account.

(2) There is no division of guest groups, and it is weak in providing further reference for decision-making.

(3) The forecasting accuracy is not high and the error is large.

Based on the above analysis, this paper establishes a hotel group division and LSTM deep learning model to forecasting different types of guest flow on the review data set of 1493 luxury hotels in Europe from 2015 to 2017. This method can effectively avoid the inaccurate forecasting caused by the sparse data of a single hotel, and shows high performance on non-stationary time series, which can guide the hotel's further decision-making.

3 HOTEL REVIEW DATA ANALYSIS AND PROCESSING

The data set used in this paper comes from Kaggle (https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe), which contains the scores of 1493 luxury hotels in Europe and 515738 guests reviews from 2015 to 2017.

The original data set contains 17 attributes: Hotel_Address,Review_Date,Average_Score,Hotel_Name,Reviewer_Nationality,Negative_Review,Review_Total_Negative_Word_Counts,Positive_Review,Review_Total_Positive_Word_Counts,Reviewer_Score,Total_Number_of_Reviews_Reviewer_Has_Given,Total_Number_of_Reviews,Tags,days_since_review,Additional_Number_of_Scoring,lat(Latitude of the hotel), lng(longitude of the hotel). Among them, a guest may have both positive and negative reviews on the hotel. Positive and negative reviews belong to two attributes of a review, which are optional; Tags contains information about the guest's accommodation, which is a variable-length string array. For example,('Leisure trip', 'Couple',' Duplex Double Room', 'Stayed 6 nights') indicates that the guest is a leisure traveling couple and stayed in duplex double room for 6 nights.

According to the goal of demand forecasting, that is, to forecast the monthly flow of leisure and business guests in the hotel group (it is worth noting that this paper mainly divides hotels into different groups based on location), weselect Review_Date, Hotel_Address, Hotel_Name, lng, lat and Tags from data set. We need to process the original data into two new data sets, one is the hotel location data set, which contains the hotel name, address, longitude and latitude attributes; the other is the time series data of a single hotel monthly guest flow from 2015 to 2017, including the hotel name, time, the number of leisure guests and the number of business guests, where the guest type needs to be separated from the Tags attribute. Some data of the data set 1 and data set 2 are shown in Tab. 1 and Tab. 2.

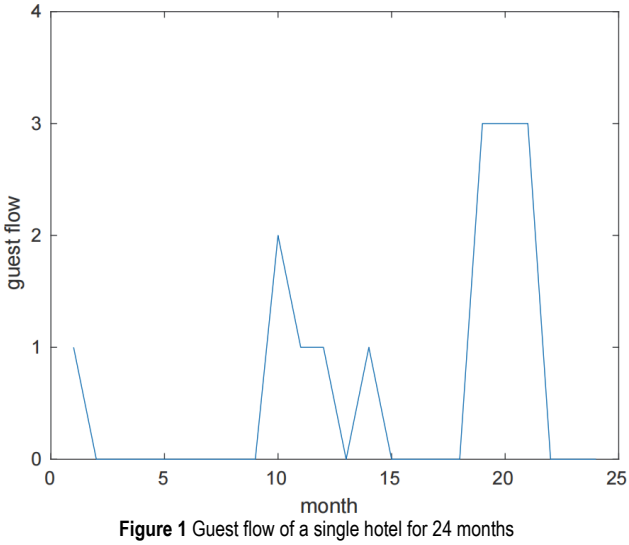

Figure 1 Guest flow of a single hotel for 24 months

Reviews reflect the guest's satisfaction with hotel services, environment, etc., and can guide the development direction of the hotel. Reviews are optionally filled in according to the personal wishes. Not all guests will leave

a review. In the case of single hotels, the review data is sparse. Fig. 1 shows the review data of a hotel for 24 months. In the figure, the number of reviews is 0 most of the time, which can effectively simulate the situation that some hotels have few guests. In fact, it is difficult for us to dig out the rules of hotel guest flow changes from such a set of sparse data. Demand forecasting using existing methods will inevitably have low forecasting accuracy and unstable results. We can effectively avoid the problem of data sparsity and improve the forecasting performance by clustering hotels with similar features, that is, dividing hotel groups.

4 METHODOLOGY

The time series method for hotel demand forecasting proposed in this paper mainly includes two steps: based on the address, longitude and latitude features of each hotel, K-means clustering algorithm is used to divide the group; the leisure and business guest flow is calculated as two parameters, and the results corresponding to the hotel group in the forecast month are obtained through the LSTM model. In this section, we will introduce the proposed demand forecasting method in detail.

4.1 Group Division

The specific group division method is as follows: first, input the identifiable address, longitude and latitude of each hotel into the K-means model for clustering, forming multiple different hotel clusters, each hotel cluster containing multiple hotels; next, the corresponding hotel cluster is regarded as a hotel group; finally, the hotel group and all hotels in the corresponding hotel group are formed into a list.

In the process of group division in this paper, K-means algorithm is used to cluster hotels with similar features to form multiple groups. K-means is a widely used unsupervised algorithm for creating data sets [28]. For a given sample set, K-means divides the samples into K clusters according to the distance between the samples, making the sample points in each cluster as close as possible, and the distance between each cluster as large as possible. We use Euclidean distance to calculate the distance between sample points in space:

$$\mathrm{dis}(x,\ y)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2} \tag{1}$$

where, n represents the spatial dimension. Suppose the input sample set $D = \{x_1, x_2, \ldots, x_N\}$, N is the number of samples, the number of clusters is K, the set of clusters $C = \{C_1, C_2, \ldots, C_K\}$, the set of cluster centroids $\mu = \{\mu_1, \mu_2, \ldots, \mu_K\}$, where:

$$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x \tag{2}$$

The equation for determining the optimal K value is as follows:

Table 1 Hotel location data set (data set 1)

Hotel_Name	Hotel_Address	lat	lng
Hotel Arena	s Gravesandestraat 55 Oost 1092 AA Amsterdam Netherlands	52.3605759	4.9159683
K K Hotel George	1 15 Templeton Place Earl s Court Kensington and Chelsea London SW5 9NB United Kingdom	51.4918878	−0.1949706
Apex Temple Court Hotel	1 2 Serjeant s Inn Fleet Street City of London London EC4Y 1LL United Kingdom	51.5137335	−0.1087512

Table 2 Hotel guest flow data set (data set 2)

Date	Hotel_Name	Leisure_Trip	Business_Trip
April 2017	222 Marylebone Road Westminster Borough London NW1 6JQ United Kingdom	9	3
May 2017	222 Marylebone Road Westminster Borough London NW1 6JQ United Kingdom	19	0
June 2017	222 Marylebone Road Westminster Borough London NW1 6JQ United Kingdom	4	3

Algorithm 1 K-means

Input: D, K
Output: C

Initialize K cluster centroids μ randomly from D;
repeat
  for $i \leq K$ **do**
    set $C_i = \emptyset$;
    ++ i;
  end for
  for $j \leq N$ **do**
    set cluster category of x_j: $\lambda_j = \arg\min_i \|x_j - \mu_i\|_2$;
    update $C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$;
    ++ j;
  end for
  for $k \leq K$ **do**
    update $\mu_k = \frac{1}{|C_k|}\sum_{x\in C_k} x$;
    ++ k;
  end for
until convergence
return C

Figure 2 The process of K-means

The process of K-means algorithm is shown in Fig. 2. Taking hotel group division as an example, the algorithm first randomly selects K samples from the input hotel sample set as the initial centroid, next sets each cluster to an empty set, and then calculates the distance from each sample to each centroid vector and adds the sample to the nearest cluster; the centroid of each cluster is finally updated. The steps except for randomly selecting the centroid are repeated until convergence.

As we all know, K-means needs to specify the number of clusters in advance, namely K, the quality of the clustering result is affected greatly by K. How to choose an appropriate K value has always been the focus and difficulty of K-means. In this paper, we select the best K value according to the elbow method [29, 30]. Compared with the method of setting the number of hotels in groups and then getting the K value by calculation in literature [10], our method can get better clustering results.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} \left\| x - \varphi_i \right\|_2 \quad (3)$$

where, *SSE* represents the clustering error of all samples, that is, the sum of the squares of the distance from each sample point to the centroid of the cluster to which it belongs, and represents the quality of the clustering result. Typically, *SSE* decreases with the increase of *K*, and tends to be stable after reaching an inflection point. The *K* value corresponding to the inflection point is considered to be the best.

4.2 Demand Forecasting

The specific demand forecasting method is as follows: first, add the historical guest flow of hotels in the same hotel group to obtain the historical guest flow of the hotel group, note that the flow of leisure and business guests is calculated separately here; and then the types of guest flow of the hotel group are input into the LSTM model to obtain the result in the forecasting month.

Hotel guest flow is affected by factors such as weather, economy, emergencies, etc. It is a kind of time series data with complex nonlinear features [7]. Due to the advantages of LSTM in processing non-stationary time series data, this paper builds an LSTM model to forecast the demand of hotel groups. LSTM is a variant of RNN with the additional function of memorizing data sequences. It remembers the early trends of data through gates and storage lines [31]. Compared with RNN, it can better capture long-term dependencies, which is useful for forecasting the long-term demand. The key of LSTM is the cell state, which is similar to a conveyor belt. LSTM contains three types of gates, namely input gate, forget gate and output gate. These gates control the selective passage of information to protect and control the cell state.

The LSTM time series forecasting model constructed in this paper includes an input layer, an output layer and two hidden layers. The parameter batch size of the input layer is 128, input dim is 1, and time step is 1; the activation function of the hidden layer is ReLU function, the gate activation function is sigmoid , and the number of hidden layer neurons is 150; the output layer selects tanh as activation function, LossL1 is used as the loss function, and the dimension of the output result is 1. In order to solve the over-fitting problem of the model on the data set, we use the Dropout algorithm [32] to improve the generalization ability of the model. This algorithm can make the activation values of certain hidden neurons stop working with a certain probability in the process of forward propagation, reduce the model's dependence on local features. In addition, gradient descent algorithm is used for the optimization of the model, and Adam algorithm is used for parameter iteration.

5 EXPERIMENTS

In the experiment, we mainly complete two tasks. One is to use the K-means model to cluster hotels and form hotel groups; the other is to build LSTM and benchmark models to forecast the monthly guest flow of a hotel group and single hotels respectively, so as to make a comparative analysis of the forecasting performance.

5.1 Data Processing

The original data set used in this paper comes from the data of 1493 luxury hotels in Europe (Kaggle). We process the original data set to obtain the geographic location data set and the monthly guest flow data set of a single hotel. The content of the data set has been introduced in the section 3.

Since hotel groups are divided according to hotel address, longitude and latitude in this paper, we need to remove the Hotel_Name attribute in data set 1 before entering the data into the K-means model. In addition, the address is converted into a numeric type through WEKA, and the data is normalized to unify the dimensions.

After getting the clustering results, we can know the group of each hotel. We count and obtain the monthly guest flow of the same hotel group from August 2015 to July 2017. After that, January 2017 is regarded as the time point to split the training set and the test set, that is, take 70% as the training data and 30% as the test data to forecast the guest flow of the hotel group.

5.2 Model Establishment

The models we build in the experiment include K-means, LSTM and benchmark models include ANN and SVR. Among them, ANN is widely used in time series forecasting, and SVR is one of the most representative nonlinear forecasting algorithms. The introduction of these two benchmark models proves that the deep learning models have more advantages than the traditional machine learning models in dealing with complex non-stationary time series. We build the above models on WEKA 3.8.5 platform of Windows 10 system. The key parameter K of K-means is determined by the elbow method; the ANN model uses a single hidden layer network structure; the SVR model uses the RBFas the kernel function, the parameter c is set to 0.15, and the parameter gamma is set to 0.02. The initial learning rate of all models is 0.001, the batch size is 128, and other parameters remain default.

5.3 Experimental Process and Result Analysis

First, we establish a K-means model to divide hotel groups, that is, to cluster hotels through certain features. The key parameter *K* of the K-means model is determined by the elbow method. In the experiment, we increase the *K* value from 2 to 30, and calculate the corresponding SSE value, as shown in Fig. 3. It can be seen that when K is 13 , the SSE value plummets to a lower level, and then gradually stabilizes. From this we can determine that setting the K value to 13 is the best choice for this paper.

Therefore, this paper uses the K-means model to cluster the hotels into 13 clusters. The cluster which the hotel belongs to is related to three features: longitude, latitude and address. The clustering results are shown in Fig. 4. 1493 hotels in Europe are divided into 13 hotel groups according to location information. The number of hotels in each hotel group is not fixed. Tab. 3 shows the number of hotels in each hotel group.

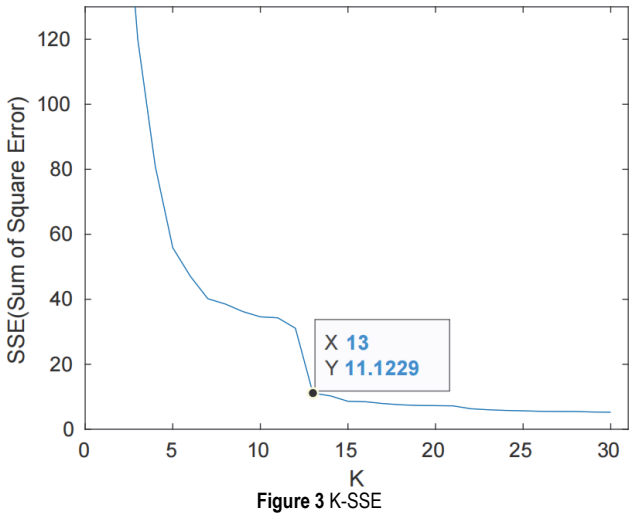

Figure 3 K-SSE

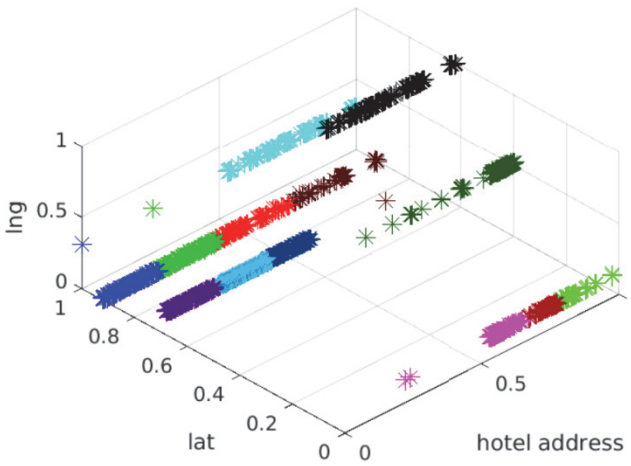

Figure 4 Clustering result

Table 3 The result of hotel group division

Hotel Group	1	2	3	4	5	6	7	8	9	10	11	12	13
Number	103	51	61	92	162	150	133	160	92	152	55	135	147

After group division, we obtain the guest flow data of each hotel group by calculation, and randomly select one of the hotel groups for demand forecasting. We build LSTM, ANN and SVR models on 70% of the data set, get 16-month forecasting results, and then test the models on 30% of the data set, get 7-month forecasting results. Here we adopt one-step forecasting. August 2015 is the first month in the data set, there is no historical data as the basis for its forecasting, so we cannot forecast this month here. Fig. 5 to Fig. 7 show the fitting curve of each model on the data set. From the figures, we can see that the number of business guests in this hotel group is far less than the number of leisure guests, and the flow of different types of guests will show peaks and lows in some months, which points out the direction for the hotel to adjust its business strategy for different types of guests in different time periods. Comparing the fitting curves of different models on the data set, we can see that LSTM has the best fit degree, followed by ANN, and finally SVR.

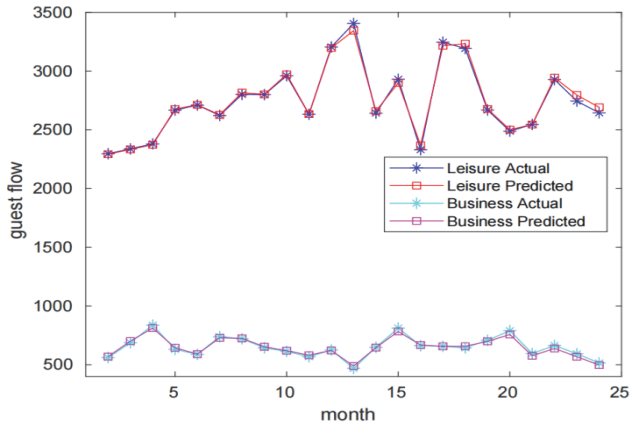

Figure 5 Fitting curve of LSTM model on the data set

In order to further compare the performance, we introduce MAE, MAPE and RMSE evaluation metrics in this paper. These three metrics are often used to evaluate the difference between the predicted and the real value in hotel demand forecasting [33]. Their value reflects the performance of the three models. The smaller the error, the better the performance of model. Tab. 4 shows the metric scores of the three models in the training and test stage, in which the best scores of each metric are shown in bold.

It can be seen from Tab. 4 that LSTM has the best scores in 5 items, ANN has the best scores in 2 items, and SVR has the worst scores. Although the score gap between the training and the test set of the LSTM model is larger than that of the other two models, its performance is basically better than them. Here we can see that the LSTM model has more advantages than ANN and SVR in forecasting hotel group guest flow.

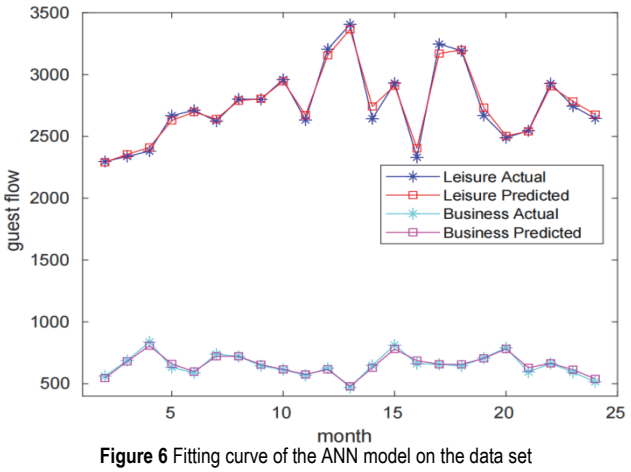

Figure 6 Fitting curve of the ANN model on the data set

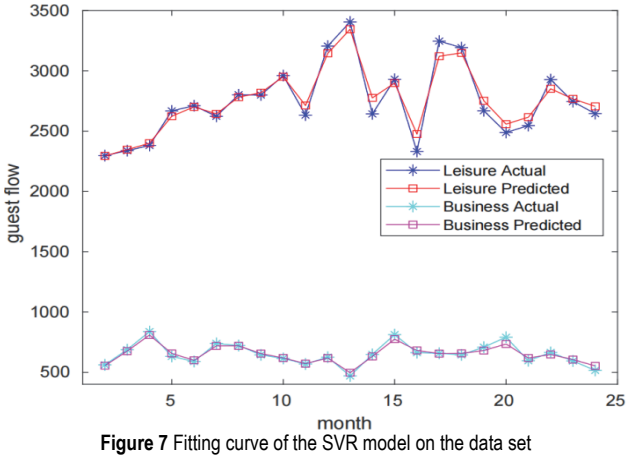

Figure 7 Fitting curve of the SVR model on the data set

Table 4 The evaluation metric scores of the three models on the data set

model	MAE		MAPE / %		RMSE	
	training	test	training	test	training	test
LSTM	**9.640**	**20.394**	**0.742**	1.894	**12.766**	**24.836**
ANN	20.678	23.548	1.396	**1.770**	27.026	**24.836**
SVR	21.486	29.151	1.421	2.039	30.367	33.522

In order to explore whether the hotel division of the group can effectively improve the accuracy of demand forecasting, and avoid the poor performance caused by sparse data, we randomly select three hotels in this hotel group (the hotel group selected for forecasting before) and use the same structure of the LSTM model to forecast the guest flow. The hotel's data is also divided into 70% training data and 30% test data. We compare the forecasting results of these single hotels and their corresponding hotel group. MAE and RMSE are affected by the data value, for different data sets; the larger the value is, the larger scores of the two metrics will be, so it is unreasonable to use them for evaluation here. Therefore, we use MAPE as an evaluation metric. Tab. 5 shows the metrics scores, where the best scores are shown in bold.

It can be seen from Tab. 5 that compared to any single hotel, the MAPE scores of the hotel group are the best. Moreover, since the data is sparse, LSTM shows a certain over-fit in the forecasting of single hotel 1, it has poor performance in test set. Therefore, the accuracy of demand forecasting for hotel groups is higher than that of a single hotel with sparse data, and group division effectively improves the forecasting performance.

Table 5 The MAPE scores of LSTM for different hotel forecasting

hotel	MAPE / %	
	training	test
hotel group	**0.742**	**1.894**
single hotel 1	5.579	44.463
single hotel 2	4.980	19.563
single hotel 3	14.683	39.992

Through the above experiments, we can draw two conclusions: (1) LSTM has more advantages than other models in terms of hotel group guest flow forecasting; (2) demand forecasting based on K-means group division can avoid the forecasting difficulties, instability and precision decline caused by sparse data of a single hotel, and can effectively improve the forecasting performance.

6 CONCLUSION

Accuracy is the key to hotel demand forecasting. Aiming at the problem of sparse historical data of a single hotel, this paper proposes a hotel demand forecasting method based on K-means group division and LSTM time series model. Taking 1493 hotels in Europe as an example, we construct a K-Means model to divide hotels into groups according to their geographic locations, and use the LSTM model to forecast the flow of different types of guests. In the experimental stage, based on the hotel's 515738 review data, we first use the elbow method to select the best *K* value to establish a K-means model, and divide the hotel into 13 groups; then, we build LSTM and its benchmark models include ANN and SVR, and introduce MAE, MAPE, RMSE to compare and evaluate the forecasting performance of the models, proving the advantage of LSTM in forecasting nonlinear time series; in addition, the LSTM model is used to predict the guest flow of the hotel group and single hotels respectively, which proves that K-means group division method can avoid problems such as inaccurate and unstable forecasting caused by sparse data. Comprehensively analyzing the experimental results, the method proposed in this paper effectively improves the forecasting performance under the condition of sparse hotel historical data, and can provide an accurate and reasonable reference for the hotel to adjust the operation decision according to guest types.

7 REFERENCES

[1] Lim, C., Chang, C., & McAleer, M. (2009). Forecasting h(m)otel guest nights in New Zealand. *International Journal of Hospitality Management*, *28*(2), 228-235. https://doi.org/10.1016/j.ijhm.2008.08.001

[2] Weatherford, L. & Kimes, S. (2003). A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting*, *19*(3), 401-415. https://doi.org/10.1016/s0169-2070(02)00011-0

[3] Croes, R. & Semrad, K. (2012). Discounting Works in the Hotel Industry: A Structural Approach to Understanding Why. *Tourism Economics*, *18*(4), 769-779. https://doi.org/10.5367/te.2012.0138

[4] Zheng, T. (2014). What caused the decrease in RevPAR during the recession? *International Journal of Contemporary Hospitality Management*, *26*(8), 1225-1242. https://doi.org/10.1108/ijchm-05-2013-0192

[5] Corgel, J., Lane, J., & Walls, A. (2013). How currency exchange rates affect the demand for U.S. hotel rooms. *International Journal of Hospitality Management*, *35*, 78-88. https://doi.org/10.1016/j.ijhm.2013.04.014

[6] Guizzardi, A. & Stacchini, A. (2015). Real-time forecasting regional tourism with business sentiment surveys. *Tourism Management*, *47*, 213-223. https://doi.org/10.1016/j.tourman.2014.09.022

[7] Zhang, B., Pu, Y., Wang, Y., & Li, J. (2019). Forecasting Hotel Accommodation Demand Based on LSTM Model Incorporating Internet Search Index. *Sustainability*, *11*(17), 4708. https://doi.org/10.3390/su11174708

[8] Aliyev, R., Salehi, S., & Aliyev, R. (2019). Development of Fuzzy Time Series Model for Hotel Occupancy Forecasting. *Sustainability*, *11*(3), 793. https://doi.org/10.3390/su11030793

[9] Zhang, Y. (2019). Research Summary of Demand Forecasting Based on Hotel Revenue Management. *Sci-tech Innovation & Productivity*, *7*, 7-12.

[10] Wu, L., Hui, Y., & Zhao, H. (2016). A Hotel Group Division and Demand Forecasting Method. *CN*. Retrieved 7 December 2016.

[11] Lee, M. (2018). Modeling and forecasting hotel room demand based on advance booking information. *Tourism Management*, *66*, 62-71. https://doi.org/10.1016/j.tourman.2017.11.004

[12] Zhang, Y. (2019). *Forecasting Hotel Demand Using Machine Learning Approaches* (Degree of Master). Cornell University.

[13] Pereira, L. (2016). An introduction to helpful forecasting methods for hotel revenue management. *International Journal of Hospitality Management*, *58*, 13-23. https://doi.org/10.1016/j.ijhm.2016.07.003

[14] Andrew, W., Cranage, D., & Lee, C. (1990). Forecasting Hotel Occupancy Rates with Time Series Models: An Empirical Analysis. *Hospitality Research Journal*, *14*(2), 173-182. https://doi.org/10.1177/109634809001400219

[15] Zhang, G., Wu, J., Pan, B., Li, J., Ma, M., Zhang, M., & Wang, J. (2017). Improving daily occupancy forecasting accuracy for hotels based on EEMD-ARIMA model. *Tourism Economics*, *23*(7), 1496-1514. https://doi.org/10.1177/1354816617706852

[16] Aydemir, E. & Gulsecen, S. (2019). Arranging Bus Behaviour by Finding the Best Prediction Model with Artificial Neural Networks. *Tehnicki vjesnik-Technical Gazette*, *26*(4), 885-892. https://doi.org/10.17559/TV-20170629201111

[17] Li, Y., Xu, M., Wen, X., & Guo, D. (2020). The Role of Internet Search Index for Tourist Volume Prediction Based on GDFM Model. *Tehnicki vjesnik-Technical Gazette*, *27*(2), 576-582. https://doi.org/10.17559/TV-20191231071057

[18] Lorenc, A., Kuźnar, M., Lerher, T., & Szkoda, M. (2020). Predicting the Probability of Cargo Theft for Individual Cases in Railway Transport. *Tehnicki vjesnik-Technical Gazette*, *27*(3), 773-780. https://doi.org/10.17559/TV-20190320194915

[19] Qian, Y., Zeng, J., Zhang, S., Xu, D., & Wei, X. (2020). Short-Term Traffic Prediction Based on Genetic Algorithm Improved Neural Network. *Tehnicki vjesnik-Technical Gazette*, *27*(4), 1270-1276. https://doi.org/10.17559/TV-20180402112949

[20] William, C. & Payares, F. (2016). A machine learning model for occupancy rates and demand forecasting in the hospitality industry. 201-211. https://doi.org/10.1007/978-3-319-47955-2_17

[21] Tsang, W. & Benoit, D. (2020). Gaussian processes for daily demand prediction in tourism planning. *Journal of Forecasting*, *39*(3), 551-568. https://doi.org/10.1002/for.2644

[22] Zhang, D. (2017). High-speed Train Control System Big Data Analysis Based on the Fuzzy RDF model and Uncertain Reasoning. *International Journal of Computers Communications & Control*, *12*(4), 577. https://doi.org/10.15837/ijccc.2017.4.2914

[23] Wang, T. (2021). An Intelligent Passenger Flow Prediction Method for Pricing Strategy and Hotel Operations. Complexity, 2021, 1-11. https://doi.org/10.1155/2021/5520223

[24] Athanasius, Z., Neamat, E., & Amir, F. A. (2008). A comparative study of the pickup method and its variations using a simulated hotel reservation data. *ICGST international journal on artificial intelligence and machine learning*, *8*, 15-21.

[25] Tse, T. & Poon, Y. (2015). Analyzing the Use of an Advance Booking Curve in Forecasting Hotel Reservations. *Journal Of Travel & Tourism Marketing*, *32*(7), 852-869. https://doi.org/10.1080/10548408.2015.1063826

[26] De Gooijer, J. & Hyndman, R. (2006). 25 years of time series forecasting. *International Journal Of Forecasting*, *22*(3), 443-473. https://doi.org/10.1016/j.ijforecast.2006.01.001

[27] Rajopadhye, M., Ben Ghalia, M., Wang, P., Baker, T., & Eister, C. (2001). Forecasting uncertain hotel room demand. *Information Sciences*, *132*(1-4), 1-11. https://doi.org/10.1016/s0020-0255(00)00082-7

[28] Casteleiro-Roca, J., Gómez-González, J., Calvo-Rolle, J., Jove, E., Quintián, H., Gonzalez Diaz, B., & Mendez Perez, J. (2019). Short-Term Energy Demand Forecast in Hotels Using Hybrid Intelligent Modeling. *Sensors*, *19*(11), 2485. https://doi.org/10.3390/s19112485.

[29] Su, T. & Dy, J. (2004). A deterministic method for initializing k-means clustering. *In 16th IEEE International Conference on Tools with Artificial Intelligence*, 784-786. IEEE.

[30] Pham, D., Dimov, S., & Nguyen, C. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal Of Mechanical Engineering Science*, *219*(1), 103-119. https://doi.org/10.1243/095440605X8298

[31] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. *In 2018 17th IEEE International Conference on Machine Learning and Applications* (*ICMLA*), 1394-1401.

[32] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97. https://doi.org/10.1109/msp.2012.2205597

[33] Wu, D., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, *29*(1), 507-529. https://doi.org/10.1108/IJCHM-05-2015-0249

Contact information:

Tianyang WANG
(Corresponding author)
City University of Macau,
Macau SAR, China
E-mail: t20091100208@cityu.mo