

# Regression Forecast of Hong Kong Passenger Flow Based on Baidu Index

Tianhao Hua \*

School of Mathematical Sciences, University of Nottingham, Nottingham, United Kingdom

\* Corresponding author: smyth3@nottingham.ac.uk

**Abstract.** Accurate forecasting of tourist arrivals is crucial for rational planning and resource allocation in the urban tourism industry. In recent years, scholars have found that web search data is correlated with tourism demand, providing new opportunities for such forecasting. This study aims to improve the accuracy of forecasting Hong Kong tourist arrivals by utilizing web search data and advanced machine learning methods. First, this paper employs large-scale web crawling techniques to collect text data related to Hong Kong tourism and applies the Latent Dirichlet Allocation (LDA) model to extract 15 key keywords. Subsequently, based on the search popularity data of these keywords on the Baidu Index, the study employs the ARIMA model and the Elastic Net Regression algorithm respectively for tourist arrival forecasting and compares the performance of the two models. The results show that compared to the traditional ARIMA model, the Elastic Net Regression performs better on multiple key indicators: 1) The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) of forecasting total visitor arrivals and overnight visitors are significantly smaller for the Elastic Net Regression; 2) The coefficient of determination ( $R^2$ ) of the Elastic Net Regression is also substantially higher than the ARIMA model, indicating stronger data fitting capability; 3) The Elastic Net Regression maintains a small error gap between the training and test sets, suggesting good generalization performance without obvious overfitting. Overall, the method of using web search data to assist tourist arrival forecasting outperforms traditional time series analysis and can improve forecasting accuracy.

**Keywords:** Automated Web Information Crawling, Latent Dirichlet Allocation, ARIMA, Elastic Net Regression, Keyword Extraction.

## 1. Introduction

The tourism industry of a city thrives on the influx of tourists, as a small number signifies a downturn in overall tourism and idle play facilities leading to decreased tourism income. However, an excessive number of tourists can also have negative impacts, such as traffic congestion when tourist numbers exceed the carrying capacity of scenic spots or cities, resulting in a decline in the overall sightseeing experience. Therefore, accurately estimating upcoming tourist volumes is crucial for cities or scenic spots. Precise volume estimates enable effective planning of tourist projects, proper guidance for visitors, and avoidance of congestion to maximize the benefits while ensuring an optimal visitor experience and achieving mutual success. With the widespread use of the Internet, more people rely on search engines to formulate their travel strategies by searching for relevant keywords before embarking on their journeys. These search engines collect data that reflects changing search popularity over time (e.g., Baidu index or Google Trends). Numerous studies have been conducted to predict tourist volumes accurately [1,2], revealing a strong correlation between search popularity and actual visitor numbers in a city; increased search behavior often predicts an increase in future tourists [3]. This correlation provides us with potential predictive tools by analyzing search popularity data to forecast future traffic patterns. However, most current studies only utilize the search popularity associated with specific keywords like city or scenic spot names as independent variables for predicting tourist numbers without fully exploring the potential applications of search popularity [4].

## 2. Exploratory Data Analysis

### 2.1. Data Acquisition and Description

Part of data acquisition, including direct website acquisition and crawler batch crawling. In this paper, the monthly number of visitors to Hong Kong for six years from 2013 to 2019 was obtained from Hong Kong Tourism Network as the dependent variable, which can also be divided into the number of overnight tourists and the total number of tourists.

In order to extract keywords related to Hong Kong tourism from the Internet, this paper collects text information related to Hong Kong tourism from the website of Ctrip through network data collection technology, including 500 link information, a total of about 4.6 million words of Chinese content.

After keyword extraction with a potential Dirichlet distribution model, about 15 strongly related keywords were obtained from the extracted text information, and the daily time series data of search popularity of these keywords from 2013 to 2019 was obtained from the Baidu Index as an independent variable, which was also divided into mobile and client search popularity.

### 2.2. Data Preprocessing and Feature Engineering

Missing value and outlier check: This paper conducts missing value and outlier checks on independent variable and dependent variable data, and no abnormal situation is found.

Distribution check: This paper conducts a normal distribution test on all data and finds that all data are roughly in line with normal distribution.

Time scale alignment: Because the data obtained from the Baidu Index is daily scale data, the data in one month are added together to obtain the monthly search popularity data of keywords, which is aligned with the dependent variable.

## 3. Method

### 3.1. Latent Dirichlet Allocation Algorithm (LDA)

Latent Dirichlet Allocation (LDA) is a generative statistical model for topic modeling designed to extract underlying topic information from a large number of documents. The basic idea of LDA is that each document is a mixture of several topics, and each topic is composed of several lexical distributions. This modeling approach can effectively capture the underlying structure in text data [5].

The core of LDA lies in the assumption that each document is generated by multiple topics, which are distributed in a certain proportion throughout the document [6]. Specifically, this article builds the model through the following steps: Select the number of documents  $M$ . Select a topic distribution for each document  $d$ : Sample from a Dirichlet distribution to get  $\theta_d$ . For each word  $w_{d,n}$ : Select a topic  $z_{d,n}$  from the topic distribution. Sample a word  $w_{d,n}$  from the vocabulary based on the lexical distribution of the selected topic.

The topic distribution  $\theta_d$  in each document  $d$  is sampled from a Dirichlet distribution of argument  $\alpha$ :  $\theta_d \sim \text{Dirichlet}(\alpha)$

The word distribution  $\phi_k$  for each topic  $k$  is also sampled from a Dirichlet distribution:  $\phi_k \sim \text{Dirichlet}(\beta)$

For the  $n$  word  $w_{d,n}$  in each document  $d$ , this paper is sampled according to the topics  $z_{d,n}$ :  $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$

#### 3.1.1. Data preprocessing

Before applying LDA, the text data should be preprocessed to some extent, including removing the stop words, such as "", " and other words with no actual information; Stem extraction or reduction: the grouping of different forms of words into one root; Word segmentation: The division of text into lexical units.

### 3.1.2. Training LDA model

The pre-processed documents are input into the LDA model for training, and the word distribution of each topic  $\phi_k$  and the topic distribution of each document  $\theta_d$  are obtained. Variable DB Bayesian inference or Gibbs sampling algorithms are often used to estimate hidden variables during training.

### 3.1.3. Topic and keyword extraction

Once the model training is completed, keywords can be extracted. For each topic  $k$ , keywords can be extracted according to the high-frequency words in  $\phi_k$ . Generally, the first  $N$  words with the highest probability are selected as the keywords of each topic.

## 3.2. Autoregressive Integral Moving Average MModel

ARIMA (AutoRegressive Integrated Moving Average) is a widely used statistical model for time series data analysis and prediction. It combines autoregressive (AR), differential (I), and moving average (MA) components to model and predict data series based on time series [7,8]. The structure of the ARIMA model can be expressed by the following formula:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} \quad (1)$$

$y_t$ : Value of the current point in time.

$B_0$ : Constant term (mean).

$B$ : autoregressive coefficient, indicating the influence of past observations.

$\theta$ : Moving average coefficient, which represents the influence of past errors (white noise).

$\epsilon_t$ : white noise error term.

Stationarity test of data: Use the unit root test to ensure that the time series is stationary. Differential processing: If the sequence is not stationary, differential processing is usually performed until the sequence is stationary. Parameter selection: Select the appropriate  $p$  and  $q$  values, usually using autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs to help determine. Model fitting: Fit the ARIMA model with the selected parameters. Model diagnostics: Check the white noise nature of the residuals to verify that the model is a good fit. The fitted model is used to make predictions about future values.

## 3.3. Elastic Network Regression Algorithm (ENR)

Elastic Net Regression is a linear regression model that combines the advantages of Ridge Regression and Lasso Regression and is particularly suitable for processing high-dimensional data and multicollinearity problems [9]. Elastic network regression can perform variable selection and regularization at the same time to improve the predictive performance and interpretability of the model.

By introducing the combination of L1 regularization and L2 regularization, elastic network regression can automatically screen features and reduce the complexity of the model [10].

In this paper, the variables are standardized: the independent variables and dependent variables are converted into vectors and standardized to ensure that the values are distributed in the same range, which is helpful for model training. Data set partitioning: The data set is randomly divided into a training set (70%) and a test set (30%) for model training and performance evaluation. The alpha parameter is then set: the alpha value determines the proportion of L1 and L2 regularization used. 0 is ridge regression, 1 is LASSO, and a value of 0.5 indicates a combination of the two. This parameter controls the sparsity of the model and trains the elastic network model using the glmnet function. In this paper, the method of 10-fold cross-validation is used to select the best  $\lambda$ . By comparing the mean square error (MSE) under different values of  $\lambda$ , the  $\lambda$  value that minimizes MSE is found. Finally, mse and rmse were calculated to evaluate the fitting effect of the model.

## 4. Results

### 4.1. Potential Dirichlet Allocations

Through the potential Dirichlet allocation algorithm, this paper finally extracts several keywords and their strengths as shown in Table 1:

**Table 1.** Keywords and their strength extracted by LDA

| Keyword    | Strength | Keyword       | Strength |
|------------|----------|---------------|----------|
| Hongkong   | 0.13     | Mooncake      | 0.025    |
| Disneyland | 0.07     | Paradise      | 0.008    |
| Mickey     | 0.008    | Christmas     | 0.007    |
| Art        | 0.005    | Beaches       | 0.004    |
| Hotel      | 0.004    | Culture       | 0.005    |
| Exhibition | 0.004    | International | 0.004    |

In order to better show the results of the model, Fig. 1 is a word cloud map of these extracted keywords:



**Figure 1.** Word cloud map of keywords

These keywords will get their search popularity data as search terms in Baidu Index.

### 4.2. Autoregressive Integral Moving Average Model

Regression is performed in the Arima model for total passenger traffic and number of overnight passengers, and the final results are as shown in Table 2.

**Table 2.** Results of ARIMA

| Name               | MSE    | RMSE   | $R^2$  | AIC      | BIC      | Ljung-Box (p-value) |
|--------------------|--------|--------|--------|----------|----------|---------------------|
| Total arrivals     | 0.5353 | 0.7317 | 0.4582 | 198.7296 | 213.3145 | 0.3363              |
| Overnight arrivals | 0.6445 | 0.8028 | 0.3477 | 208.3753 | 218.0506 | 0.7699              |

### 4.3. Elastic Network Regression

Table 3 shows the results of elastic network regression, where the total search volume and overnight passenger traffic have smaller MSE and RMSE, and the gap is small, indicating that the model does not find obvious overfitting phenomenon.

**Table 3.** Results of Elastic net regression

|                      |              | Total ridership |        |                | Overnight passenger flow |        |                |
|----------------------|--------------|-----------------|--------|----------------|--------------------------|--------|----------------|
|                      |              | MSE             | RMSE   | R <sup>2</sup> | MSE                      | RMSE   | R <sup>2</sup> |
| Total search volume  | Training set | 0.2287          | 0.4782 | 0.7682         | 0.1969                   | 0.4438 | 0.8089         |
|                      | Test set     | 0.3555          | 0.5962 | 0.6366         | 0.3377                   | 0.5811 | 0.6213         |
| Mobile search volume | Training set | 0.3384          | 0.5818 | 0.6570         | 0.3829                   | 0.6188 | 0.6286         |
|                      | Test set     | 0.4764          | 0.6903 | 0.5130         | 0.3334                   | 0.5775 | 0.6260         |

Fig. 2 shows the results predicted by the model compared to actual observations. The color of the dots represents the data source (training or test set), and the red dashed line below indicates the line where the predicted value ideally agrees with the actual value. The top left is the regression result of total search volume and total passenger traffic, the top right is the regression result of mobile search volume and total passenger traffic, the bottom left is the regression result of total search volume and overnight passenger traffic, and the bottom right is the regression result of mobile search volume and overnight passenger traffic. The following three graphs show this pattern.

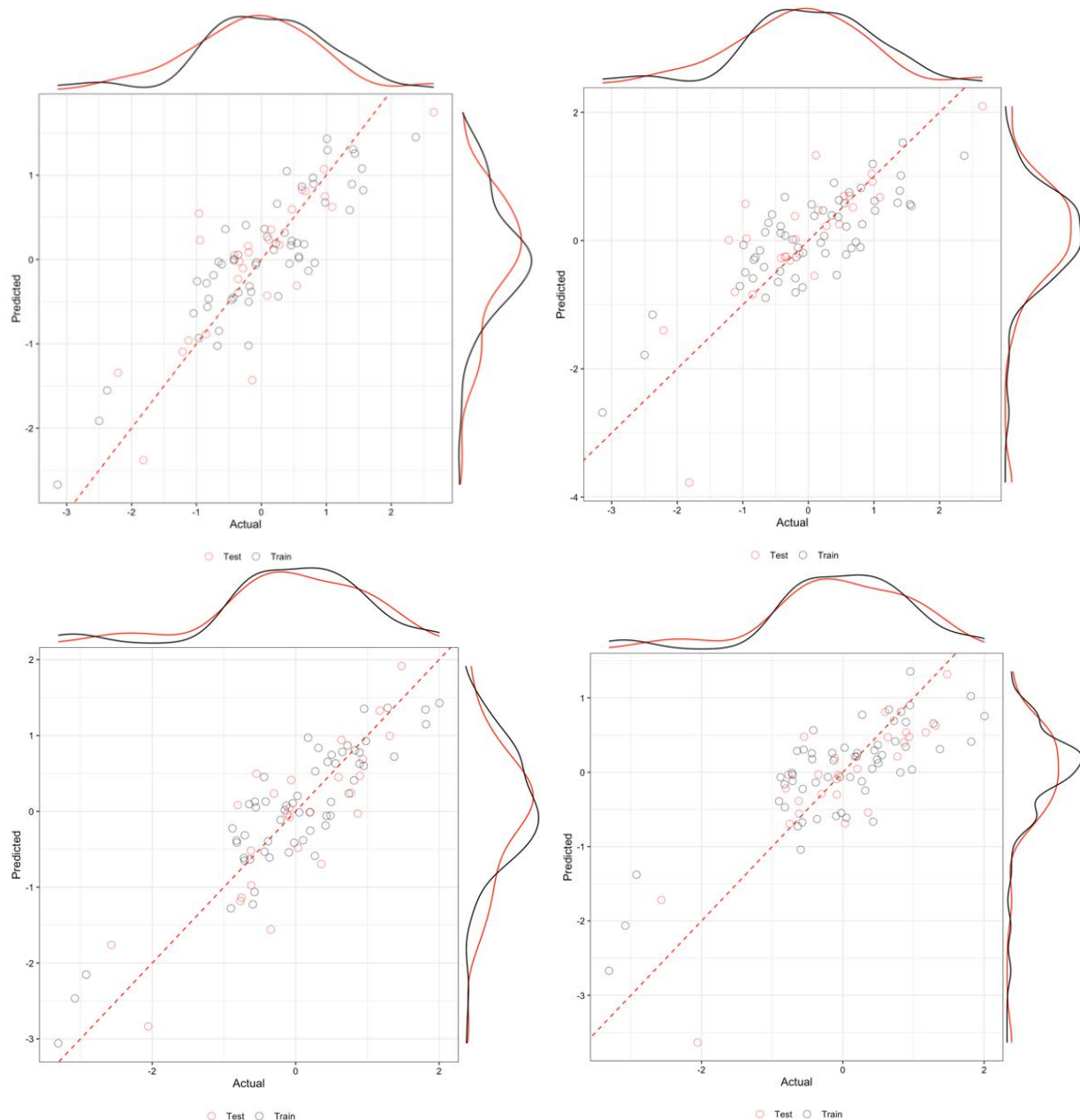


**Figure 2.** A comparison of the predicted and actual values

No significant difference was found in the scatter distribution of the training set and the test set, and the distribution was basically near the reference line, indicating that the model performance was normal and there was no obvious overfitting phenomenon.

Fig. 3 is a line chart of mean square error (MSE) for different lambda values, which shows the mean square error (MSE) of the model on the training set as the lambda values change. The X-axis is the lambda value on the logarithmic axis, while the Y-axis is the corresponding mean square error. MSE usually decreases with the increase of lambda value, indicating that moderate regularization can reduce the complexity of the model and improve its generalization ability.

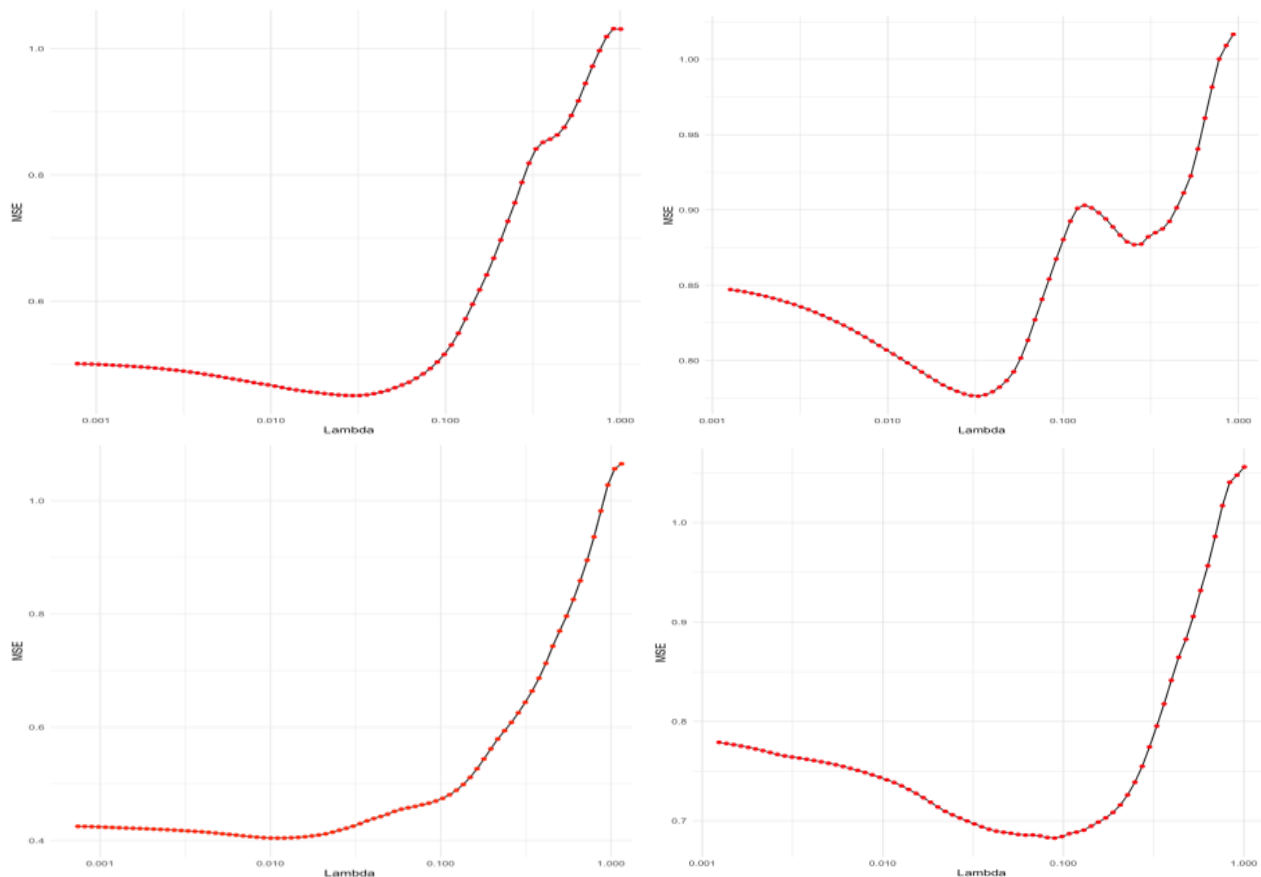


**Figure 3.** MSE line chart for different lambda values

Fig. 4 is a line chart of the final prediction, which shows the trend of actual and predicted values over the sample index on the test set. Actual values are represented by solid lines and predicted values are represented by dashed lines. Better regression results can be seen together with the above table.

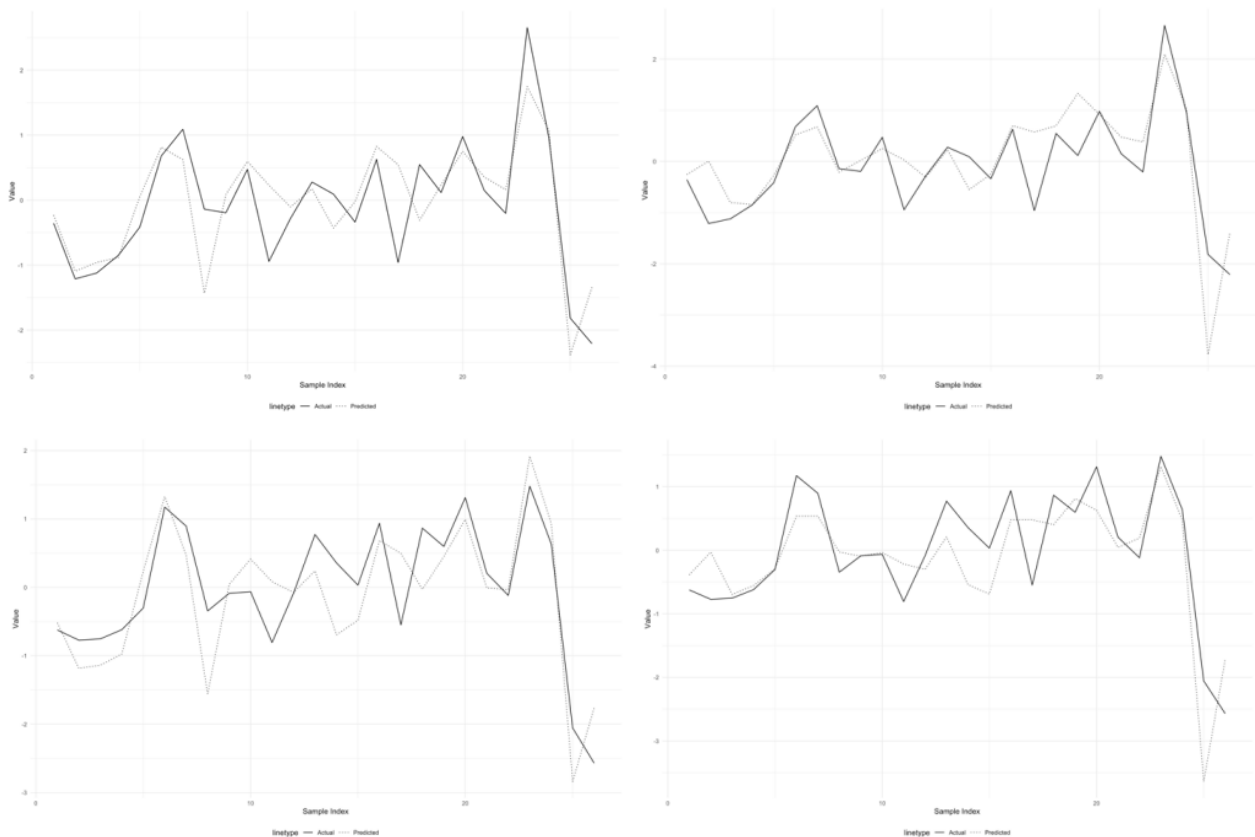


**Figure 4.** Line chart of predicted results

## 5. Discussion

Comparing the performance indicators of the ARIMA model and the elastic network regression model respectively in the data the results, the elastic network regression shows better results in several aspects.

### 5.1. Mean square Error (MSE) and Root Mean Square Error (RMSE)

In the ARIMA model, the MSE for total ridership and overnight passengers is 0.5353 and 0.6445, respectively, while the MSE for the elastic network regression model on the test set is 0.3555 and 0.3377 for overnight passengers. The MSE for the elastic network regression model on the test set is 0.3555 and 0.3377 for overnight passengers. Obviously, the error of elastic network regression model on these two prediction targets is small, indicating that its prediction accuracy is higher.

In terms of RMSE, the ARIMA model has a total ridership of 0.7317 and overnight passengers of 0.8028; The test sets RMSE of elastic network regression are 0.5962 (total passenger flow) and 0.5811 (overnight passenger flow) The test sets RMSE of Elastic network regression are 0.5962 (total passenger flow) and 0.5811 (overnight passenger flow) respectively, which also shows that the elastic network regression model has better error control ability.

### 5.2. Determination Coefficient ( $R^2$ )

The  $R^2$  value indicates the explanatory power of the model, and the closer the value is to 1, the better the model fits the data. In the ARIMA model, the  $R^2$  for total ridership and overnight passengers is 0.4582 and 0.3477, respectively, while the  $R^2$  for elastic network regression on the test set is 0.6366 (total ridership) and 0.6213 (overnight ridership), respectively. The  $R^2$  of the elastic network regression model is significantly higher than that of the ARIMA model, indicating that it has a better ability to fit the data.

In summary, elastic network regression is superior to ARIMA model in MSE, RMSE,  $R^2$  and other key indicators, and has better generalization ability. The keywords selected by this model have better progress in the prediction of overnight visitors, indicating that the search volume of keywords obtained through text analysis can help improve the progress of the prediction model. Better forecast of total ridership and overnight ridership. Comparing the performance indicators of the ARIMA model and the elastic network regression model respectively in the data results, the elastic network regression shows better results in several aspects.

### 5.3. Mean square Error (MSE) and Root Mean Square Error (RMSE)

In the ARIMA model, the MSE for total ridership and overnight passengers is 0.5353 and 0.6445, respectively, while the MSE for the elastic network regression model on the test set is 0.3555 and 0.3377 for overnight passengers. Obviously, the error of the elastic network regression model on these two prediction targets is small, indicating that its prediction accuracy is higher.

- In terms of RMSE, the ARIMA model has a total ridership of 0.7317 and overnight passengers of 0.8028; The test sets RMSE of elastic network regression are 0.5962 (total passenger flow) and 0.5811 (overnight passenger flow) respectively, which also shows that the elastic network regression model has better error control ability.

### 5.4. Determination Coefficient ( $R^2$ )

The  $R^2$  value indicates the explanatory power of the model, and the closer the value is to 1, the better the model fits the data. In the ARIMA model, the  $R^2$  for total ridership and overnight passengers is 0.4582 and 0.3477, respectively, while the  $R^2$  for elastic network regression on the test set is 0.6366 (total ridership) and 0.6213 (overnight ridership), respectively. The  $R^2$  of the elastic network regression model is significantly higher than that of the ARIMA model, indicating that it has a better ability to fit the data.

In summary, elastic network regression is superior to the ARIMA model in MSE, RMSE,  $R^2$  and other key indicators, and has better generalization ability. The keywords selected by this model have better progress in the prediction of overnight visitors, indicating that the search volume of keywords obtained through text analysis can help improve the progress of the prediction model. Better forecast of total ridership and overnight ridership.

## 6. Conclusion

In this study, the potential Dirichlet allocation (LDA) model is used to extract keywords from a large number of text information, and the search volume data of these keywords is used to assist regression prediction of passenger flow in Hong Kong. The experimental results show that the proposed method is significantly more accurate than the traditional autoregressive integrated moving average (ARIMA) model.

However, there are some limitations to this study. First, the amount of text data we use and the number of keywords extracted are relatively limited and may not fully reflect the complexity of the travel market. Second, the predictive power of the model still has room for improvement in long-term forecasting. In response to these problems, we propose the following improvement suggestions: In the actual commercial application scenario, a real-time updated large-scale keyword extraction model can be built and trained. This will help ensure the timeliness of keywords and more comprehensive coverage of all aspects of the tourism industry.

Consider adopting more advanced machine learning techniques, such as long short-term memory (LSTM) neural network models. The LSTM model performs well in processing time series data and is able to capture long-term dependencies, which is expected to further improve the accuracy of tourist volume prediction. Introduce more external factors, such as exchange rate changes, and the impact of major events (such as the Olympic Games, World Expos, etc.). These factors can have a significant

impact on travel demand, and including them in the model helps improve the accuracy and robustness of the forecast.

In conclusion, this study provides an innovative framework for using search data to predict travel demand. Despite some limitations, the results show that this approach has significant potential. Future research can focus on expanding the data size, optimizing the algorithm, and introducing more relevant factors to further improve the accuracy and practicability of the prediction. This is not only important for academic research but also provides a valuable tool for decision makers in the tourism industry to develop more accurate marketing strategies and resource allocation plans.

## References

- [1] W. Kunpeng, P. Xiao, B. Yekun, and Z. Wenyan, Modeling and Forecasting Tourism demand in Lhasa based on Baidu Index, *Journal of Xizang University*. 2017.
- [2] R. Huan, L. Ting, K. Junfeng, P. Ning, L. Minjing, and A. Shunyi, A Method for predicting the daily tourist size of Cities based on Baidu Index," *Journal of Hangzhou Normal University*. 2017.
- [3] X. Huang, L. Zhang, and Y. Ding, The Baidu Index: Uses in predicting tourism flows eA case study of the Forbidden City, *Tourism Management*, 2017, 58: 301 - 306.
- [4] L. Yao, R. Ma, and H. Wang, Baidu index-based forecast of daily tourist arrivals through rescaled range analysis, support vector regression, and autoregressive integrated moving average, *Alexandria Engineering Journal*, 2021, 60 (01): 365 - 372.
- [5] S. Tirunillai and G. J. Tellis, Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation, *Journal of Marketing Research*, 2014, 51 (04): 463 - 479.
- [6] H. Edison and H. Carcel, Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts, *Applied Economics Letters*, 2021, 28 (01): 38 - 42.
- [7] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, *Data in Brief*, 2020, 29: 105340.
- [8] F. Chang, H. Huang, A. H. Chan, S. Shing Man, Y. Gong, and H. Zhou, Capturing long-memory properties in road fatality rate series by an autoregressive fractionally integrated moving average model with generalized autoregressive conditional heteroscedasticity: A case study of Florida, the United States, 1975 – 2018, *Journal of Safety Research*, 2022, 81: 216 - 224.
- [9] H. Han and K. J. Dawson, applying elastic-net regression to identify the best models predicting changes in civic purpose during the emerging adulthood, *Journal of Adolescence*, 2021, 93 (01): 20 - 27.
- [10] C. Hans, Elastic Net Regression Modeling with the Orthant Normal Prior, *Journal of the American Statistical Association*, 2011, 106 (496): 1383 - 1393.