Machine Learning in Internet Search Query Selection for Tourism Forecasting

Xin Li, Ph.D.
Donlinks School of Economics and Management
University of Science and Technology Beijing
Beijing, China
Email: drxinli@ustb.edu.cn

Hengyun Li*, Ph.D.
Hospitality and Tourism Research Centre (HTRC)
School of Hotel and Tourism Management,
The Hong Kong Polytechnic University,
Hong Kong SAR, China
Email: neilhengyun.li@polyu.edu.hk

Bing Pan, Ph.D.
Department of Recreation, Park, and Tourism Management,
College of Health and Human Development,
The Pennsylvania State University, United States
Email: bingpan@psu.edu

Rob Law, Ph.D.
Hospitality and Tourism Research Centre (HTRC)
School of Hotel and Tourism Management,
The Hong Kong Polytechnic University,
Hong Kong SAR, China
Email: rob.law@polyu.edu.hk

* Corresponding Author: Hengyun Li

Acknowledgments
The author(s) disclose receipt of the following financial support for the research, authorship, and/or publication of this article: Research funds from the National Natural Science Foundation of China (No. 71601021), Fundamental Research Funds for the Central Universities (No. FRF-TP-19-067A1), and Hospitality and Tourism Research Centre (HTRC Grant) of the School of Hotel and Tourism Management, The Hong Kong Polytechnic University (Project Account Code: 5-ZJLT).

Machine Learning in Internet Search Query Selection for Tourism Forecasting

Abstract:

Prior studies have shown that Internet search query data have great potential to improve tourism forecasting. As such, selecting the most relevant information from large amounts of search query data is crucial to enhancing forecasting accuracy and reducing overfitting; however, such feature selection methods have not been considered in the tourism forecasting literature. This study employs four machine-learning-based feature selection methods to extract useful search query data and construct relevant econometric models. We examined the proposed methods based on monthly forecasting of tourist arrivals in Beijing, China along with weekly forecasting of hotel occupancy in the city of Charleston, South Carolina, USA. Our findings indicate that the forecasting model with the selected search keywords outperformed the benchmark ARMAX model without feature selection in forecasting tourism demand and hotel occupancy. Therefore, machine learning methods can identify the most useful search query data to significantly improve forecasting accuracy in tourism and hospitality.

Keywords: tourism forecasting; hotel occupancy; search query data; machine learning; feature selection

1. Introduction

Tourism demand forecasting plays a crucial role in the tourism and hospitality industry. Recently, accurate prediction of tourism demand has become an increasingly important research topic. Many studies to date have focused on forecasting tourism demand using diverse quantitative methods, including time series, econometric, artificial intelligence, and integrated models (Chen et al. 2019; Gunter and Önder 2015; Song, Qiu, and Park 2019; Song and Li 2008; Song and Witt 2000; Sun et al. 2016). Recent forecasting methods include deep learning (Law et al. 2019), spatial temporal models (Yang and Zhang 2019), ensemble empirical mode decomposition (Li and Law 2020), Bayesian global vector autoregressive (Assaf et al. 2019), Markov switching models (Valadkhani and O'Mahony 2018), and pooling (Long, Liu, and Song 2019). These developments reflect researchers' efforts to propose a more effective forecasting method to significantly improve the forecasting accuracy of tourism demand.

Scholars have found that Internet big data, such as search query data, can be used to predict tourism demand accurately. In particular, search query data can reflect tourist behavior and supplement traditional data sources to predict tourism demand (Choi and Varian 2012; Yang et al. 2015). To incorporate search query data into forecasting models, researchers have proposed combining one or several representative indices as explanatory variables to reduce model complexity. Keyword selection from search queries largely determines tourism forecasting performance. In the existing literature, many studies have

defined keywords on the basis of prior domain knowledge and then collected search query data from Baidu or Google to represent tourists' interests (Brynjolfsson, Geva, and Reichman 2016; Sun et al. 2019). Common approaches to index aggregation include shift and summation, principle component analysis, and the generalized dynamic factor model. For example, Sun, Wei, Tsui, and Wang (2019) aggregated one index from 22 Baidu search query data using the shift and sum method for tourism forecasting. Li, Pan, Law, and Huang (2017) extracted a composite index from 45 search query data as an explanatory variable via a generalized dynamic factor model. However, these studies did not determine which search query data are most helpful for improving tourism forecasting accuracy.

When selecting search keywords, the tradeoff between search query data coverage and accuracy has been deemed pivotal (Geva et al. 2017). Incorporating a large group of search query data into searching may reveal relevant information, but it may also introduce irrelevant noise and cause problems such as spurious correlation and data overfitting (Brynjolfsson, Geva, and Reichman 2016; Geva et al. 2017; Song and Liu 2017). Researchers need to determine which search keywords should be retained to maintain the accuracy. The question of how to automatically select the most appropriate combinations of search query data related to tourism demand from search engines that can effectively improve the forecasting accuracy remains unanswered.

This study aims to investigate whether machine learning methods can help select the most useful search query data to improve predictions of tourism demand and hotel occupancy compared to a benchmark ARMAX model without machine learning. Moreover, if machine

learning methods are well suited to tourism demand forecasting, which method realizes the best forecasting performance? To achieve these research objectives, we conducted empirical studies that include (1) the monthly tourism demand forecasting in Beijing using Baidu search data; and (2) the weekly hotel occupancy forecasting using Google data. In addition, several machine-learning-based feature selection methods (i.e., filter-based feature selection, recursive feature selection, genetic algorithm feature selection, and random forest feature selection) were applied to extract appropriate search query data subsets for incorporation into our forecasting models. Forecasting results reveal the effectiveness of machine learning feature selection methods in improving the forecasting of tourism demand and hotel occupancy.

The remainder of this paper is organized as follows. Section 2 presents a literature review. Section 3 introduces our four feature selection methods and describes the methodology. Section 4 outlines our empirical results, including the monthly Beijing tourism forecasting and the robustness of weekly hotel occupancy forecasting. The last section provides concluding remarks.

2. Literature Review

Three streams of literature are relevant to our work: (1) tourism forecasting models with search query data; (2) Internet search query data selection; and (3) the current state of research and application of machine-learning-based feature selection.

2.1 Tourism forecasting models with search query data

Tourism forecasting has become increasingly important due to an urgent need for timeliness and accuracy in forecasting tasks (Frechtling 2012; Guizzardi, and Stacchini 2015; Hassani et al. 2017; Pai, Hung, and Lin 2014; Peng, Song. and Crouch 2014; Shen, Li, and Song 2011; Zhou-Grundy and Turner 2014). Existing literature in tourism and hospitality demonstrates that Internet search query data have become an important variable to increase tourism forecasting accuracy (Li et al. 2018a; Song, Qiu, and Park 2019). Pan, Wu, and Song (2012) was one of the first to demonstrate how Google search query data could improve forecasting accuracy of hotel room demand. Scholars have continued to adopt different models to effectively analyze the search query data and to improve forecasting accuracy. Table 1 presents an overview of selected work in the tourism literature.

[Please insert Table 1 near here]

When tourists plan trips, they may refer to Internet search engines to retrieve information using keywords (Fesenmaier et al. 2011; Yang et al. 2015). Search query data from Google and Baidu are most widely used for different predicted contexts. Yang et al. (2015) indicated that Baidu search data perform better when forecasting tourist arrivals in

China; while Google search data are found to be suitable for the forecasting of countries and cities that mainly speak English (Önder 2017; Pan and Yang 2017). Search query data have been used to forecast international tourism demand of a tourism destination from multiple source markets (e.g., Li and Law 2020). Scholars also examined the predictive ability of search query data in forecasting single time series such as the total number of tourist arrivals to one destination (e.g., Hu and Song 2019; Li et al. 2017; Pan and Yang 2017; Sun et al. 2019; Yang et al. 2015). Both types of research have revealed the usefulness of search query data in tourism demand prediction.

In terms of methodologies, time series, econometric, artificial intelligence, and hybrid models have been adopted in the existing forecasting literature. Time series and econometric models have accounted for a larger proportion of studies. Search query data are incorporated into the models as an explanatory variable such as AR, ARMA, ARIMA, seasonal ARIMA, ADL, TVP, and VAR (Bokelmann and Lessmann 2019; Huang, Zhang and Ding 2017; Li and Law 2020; Önder 2017; Pan, Wu, and Song 2012; Park, Lee, and Song 2017). Advanced econometric models have also been developed for modelling search query data. For example, Bangwayo-Skeete and Skeete (2015) adopted an AR-MIDAS model that predicted monthly tourist arrivals with weekly Google search data. Camacho and Pacce (2018) proved that a dynamic factor model can improve tourism forecasting accuracy compared to an AR model. Li et al. (2017) proposed a GDFM model to analyze the search query data to improve forecasting performance.

Artificial intelligence models have been applied to forecast tourism demand with search query data. For example, Hu and Song (2019) demonstrated that a BPNN model can outperform ARIMA and ADL models in forecasting tourist arrivals from Hong Kong to Macau. Sun et al. (2019) suggested that a kernel extreme learning machine method can improve the forecasting accuracy of Beijing tourism demand compared to ARIMA models. Law et al. (2019) proved the ability of a deep learning method in forecasting Macau tourist arrivals. Moreover, hybrid models have been adopted in tourism forecasting with search query data recently. Li et al. (2018b) combined an adaptive differential evolution with the BPNN to further enhance the forecasting. Wen, Liu, and Song (2019) proposed a hybrid model to combine a linear ARIMA and a non-linear AR model to improve forecasting accuracy. Zhang et al. (2017) combined the Bat algorithm and support vector regression to improve forecasting performance.

Different models have their own advantages, and no single method can outperform in all forecasting situations (Song, Qiu, and Park 2019). Search query data are considered as explanatory variables that can influence forecasting performance. Therefore, the selection of search query data has become important since such a selection process determines not only which data are incorporated into the model, but also whether forecasting accuracy can be improved.

2.2 Internet search query data selection

When forecasting using search query data, concerns about the selection of search

query keywords have become unavoidable as search data selection greatly influences forecasting quality. Table 1 suggests that search query data are mainly selected based on two approaches: intuition and prior domain knowledge (Brynjolfsson, Geva, and Reichman 2016) and Google's search query index.

The first keyword selection method has been widely applied in the existing literature. Yang et al. (2015) and Li et al. (2017) defined several aspects of tourist activities including transportation, dining, lodging, shopping, recreation, and tours. Studies such as Li et al. (2018b) and Law et al. (2019) followed their search keywords selection frameworks but made minor adjustments by adding one aspect relevant with clothing. Hu and Song (2019) excluded search terms from shopping and tours, since these activities are not the motivation for visitors from Hong Kong to Macau. As discussed by Geva et al. (2017), the advantage of such search data selection is achieving a high coverage of search keywords; however, by selecting more search query data, the overfitting problem resulted from irrelevant data is unavoidable. It would possibly influence the accuracy of forecasting models.

Some researchers directly obtained search query data by using Google category index in Google's search engine (Camacho and Pacce 2018; Önder 2017). For example, Bangwayo-Skeete and Skeete (2015) used a search index for 'hotels and flights' in Google to predict tourist arrivals in the Caribbean. Li and Law (2020) used the Google query index from Hong Kong travel subcategories to improve out-of-sample forecasting accuracy for Hong Kong tourist arrivals. However, researchers cannot obtain the specific combination of keywords

included in the Google index, because they are not transparent to users (Choi and Varian 2012).

Therefore, selecting the most relevant data that reflect both the coverage and accuracy of search keywords is important for tourism forecasting with search query data. Existing studies have not yet considered an optimal search query data selection method to obtain an improved subset, but extracted one or several representative indices as explanatory variables in forecasting models. For instance, Li et al. (2017) extracted a composite index using the generalized dynamic factor model. Xie et al. (2020) selected principle components using a kernel principle component analysis to reflect information contained in search query data. Yang et al. (2015) aggregated search query data into an index using shift and summation. An extracted representative index can reflect useful information at an aggregated level, but it is still difficult to determine which search keywords are most relevant for improving tourism forecasting.

2.3 Machine-learning-based feature selection

Feature selection is an essential procedure that removes irrelevant information to improve algorithm performance (Chandrashekar and Sahin 2014). When considering collected search query data as a feature set, researchers can obtain a subset of data that can achieve higher performance (Domingos 2012; Guyon et al. 2002). One advantage of feature selection is avoidance of overfitting and enhanced prediction performance (Guyon and Elisseeff 2003). In this study, a machine-learning-based feature selection method is adopted

to remove redundant and extraneous information from search query data, and to retain the most relevant subset of keywords that helps achieve a considerable coverage and an improved forecasting accuracy.

Machine-learning-based feature selection can be applied to supervised learning in a forecasting context (Cai et al. 2018; Kodratoff and Michalski 2014). The criteria to keep or remove a feature depends on whether it can improve forecasting performance based on different machine learning algorithms (Cui et al. 2017; Kursa and Rudnicki 2010). Each feature selection method has its own advantages and disadvantages, and we focus here on four classical methods: filter-based feature selection, recursive feature selection, genetic algorithm feature selection, and random forest feature selection. Our review provides a brief introduction of each method.

The filter-based feature selection method is a relatively fast algorithm that is particularly useful for overcoming overfitting (Yu and Liu 2004). Features are often ranked using scores based on specific criteria, such as the correlation coefficient, Chi-squared test, and information gain (Saeys, Inza, and Larrañaga 2007). Features with low scores, as calculated by an algorithm, are removed (Chandrashekar and Sahin 2014).

Recursive feature selection selects a subset of features by recursively removing features to achieve the required maximum performance and minimum number of features (Yan and Zhang 2015). First, the importance of each feature is obtained after training an initial set of features. Second, the least important features are eliminated from the set. This selection procedure is recursively repeated until the desired subset of features is assembled

(Guyon et al. 2002).

Genetic algorithm feature selection obtains the most appropriate subset of data based on a genetic algorithm that eliminates and maintains features (Saeys, Inza, and Larrañaga 2007). The key elements of a genetic algorithm are selection, crossover, and mutation. This approach first selects individual features with high fitness values from a current generation. Crossover recombines the chromosomes of two parents to generate new individuals in the next generation. Mutation is the process of changing genes randomly selected in the current chromosome (Huang and Wang 2006). The method is computationally feasible for generating suitable results (Chandrashekar and Sahin 2014).

Random forest feature selection is a popular and efficient machine learning algorithm constructed based on regression trees. It combines many binary decision trees, built using several bootstrap samples, from a learning sample and selects a subset of explanatory variables randomly at each node (Genuer, Poggi, and Tuleau-Malot 2010; Sylvester et al. 2018). The random forest method then determines the optimal subset of features based on model aggregation of classification and regression (Breiman 2001). This approach has been adopted for variable selection in several fields, such as operations and supply chain management (Cui et al. 2017).

2.4 Research gap

In summary, prior studies have indicated that Internet search query data can enhance tourism demand forecasting performance. However, several research gaps related to the use

of search query data in tourism forecasting should be addressed. Scholars have tended to either choose keywords from search engines based on intuition and prior knowledge or use the Google query index without a definitive selection process; no prior studies have explored which keyword combinations comprehensively reflect tourism demand. Questions to be considered include whether more search keywords are better for improving the forecasting accuracy of tourism demand? The notion of how to balance search keyword coverage with the accuracy of forecasting performance remains unresolved. In other words, a rigorous procedure regarding the selection of search query data has not yet been proposed.

Specifically, machine-learning-based feature selection has not yet been applied to the selection of search query data for tourism forecasting despite being an effective method for analyzing large volumes of data. Feature selection methods can choose the most appropriate combinations of search query data to obtain better forecasting performance with higher efficiency. Feature selection has been widely incorporated to process large volumes of data in fields such as bioinformatics to realize faster and more effective models (Cai et al. 2018). In tourism, however, the application of feature selection to improve forecasting performance using search query data remains limited.

3. Methodology

The four machine-learning-based feature selection methods used in this study include: filter-based feature selection, recursive feature selection, genetic algorithm feature selection, and random forest feature selection.

Filter-based feature selection is used to select the most relevant search query data based on criteria such as the correlation coefficient [*R(i)*] and information gains [$IG(X_i, Y)$]:

$$R(i) = \frac{\text{cov}(X_i, Y)}{\sqrt{\text{var}(X_i) * \text{var}(Y)}}$$

$$IG(X_i, Y) = H(X_i) - H(X_i \mid Y)$$

where $X = \{X_1, X_2, \cdots, X_n\}$ denotes *n*-dimensional search query data, and Y represents the predicted tourist arrival data. cov(X_i,Y) is the covariance of search query data and tourist arrival data; var(Y) is the variance of tourist arrival data; and $H(X_i)$ and $H(X_i|Y)$ indicate the entropy and conditional entropy, respectively (Chandrashekar and Sahin 2014).

Recursive feature selection is an iterative procedure used to eliminate irrelevant features and retain the most appropriate ones. Each feature is ranked based on its calculated importance. At each iteration, the importance of features is measured and the top ranked features are retained. The recursive feature selection algorithm is described in Table A.1 of the Appendix.

Genetic algorithm feature selection is a general adaptive optimization search method (Chandrashekar and Sahin 2014). To select the best search data subset, we first need to set the maximum generations, population per generation, crossover, and mutation

probability. Subsets of search data can then be generated through crossover and mutation, and algorithm performance is evaluated using the fitness function (Xue, Yao, and Wu 2018). The most appropriate subset is obtained after *K*-fold cross-validation. For details about the genetic algorithm, see Das, Das, and Ghosh (2017).

Random forest feature selection is an efficient machine learning method that applies bootstrap samples to combine decision trees and randomly selects a subset of variables at each node (Cui et al. 2017). The subset is selected based on its importance, which is computed using the error in the out-of-bag (OOB) sample. The OOB sample refers to the set of observations not used to build the current trees. The random forest selection procedure is described in Table A.2 of the Appendix.

Accordingly, a forecasting framework is proposed in this paper to answer the research questions by performing four major steps: (1) search keywords selection, (2) machine-learning-based feature selection, (3) econometric modelling, and (4) forecasting evaluation. Figure 1 depicts the forecasting framework.

[Please insert Figure 1 near here]

1 In the first step, we collected search query volume data from search engines such as Google and Baidu. We proposed extracting tourism-demand-related keywords to reflect tourists' attention to various activities including dining, lodging, traffic, recreation, shopping, and tourism. These keywords reflected tourists' decisions about various aspects of a trip during their travel planning process (Yang et al. 2015). The selected keywords should be comprehensive and reflect various

dimensions of tourism demand related to tourist activities. Tourists who express their interest in a certain travel destination will likely search for “special attractions” or “special food” through a search engine. They also search for keywords such as “travel guides” or “travel plans” to obtain more information about potential destinations. It should be noted that not all keywords associated with tourism concepts can be included; search engines do not return keyword data series containing few results (Sun et al. 2019). For example, data are available for the keyword phrase “Beijing tourism”, but a query for “How to travel in Beijing” returns no results. For the purposes of this paper, a group of search query data was generated based on selected keywords.

2 In the second step, the machine-learning-based feature selection methods were used to extract subsets of search query data from the above collected data set. After the feature selection procedure, the dimensions of the search query data will be reduced but the forecasting accuracy can be improved because of the elimination of irrelevant data. Since we introduced different machine learning based methods, we can evaluate which method can effectively deal with search query data. It should be noted that the particular focus is the automatic selection from the perspective of machine learning techniques. Although the methods have different criteria to retain the search query data, the fundamental idea is to estimate if the addition of data improves or reduces forecasting accuracy. Therefore, the ‘black-box’ nature of machine learning methods simplifies the data

selection step, with no further manual examination is needed. Results of the above four feature selection methods can be obtained through R software using the "caret" and "Boruta" packages (Kursa and Rudnicki 2010).

3 In the third step, econometric models were constructed to incorporate the search query data and to predict tourism demand. We focus on the improvement of forecasting accuracy with feature selection compared to that without selection. A benchmark ARMAX model was obtained by adding all search query data on the basis of the ARMA model, without any feature selection method. To evaluate whether the machine-learning-based feature selection method improves the tourism forecasting performance, four econometric models incorporating the subsets of search query data extracted from the aforementioned four feature selection methods were then constructed, and their respective forecasting performances were compared with the benchmark ARMAX model above.

The ARMAX model is described as

$$y_t = \alpha + \sum_{i=0}^{p} \beta_i y_{t-i} + \varepsilon_t + \sum_{i=0}^{q} \lambda_i \varepsilon_{t-i} + \sum_{i=0}^{k} \gamma_i x_{t-i}$$

where y_t indicates tourist arrival data. $y_t = \alpha + \sum_{i=0}^{p} \beta_i y_{t-i} + \varepsilon_t + \sum_{i=0}^{q} \lambda_i \varepsilon_{t-i}$ is a classical ARMA model without search query data. The lag orders *p, q* and *k* are determined by the Akaike information criterion (AIC). For convenience of estimation in the ARMAX model, we took one index x_t aggregated from all search query data, without feature selection, using the shift and sum method (Yang et al. 2015).

The econometric models with the selected search query data obtained from machine-learning-based feature selection can be written as:

$$y_t = \alpha + \sum_{i=0}^{p} \beta_i y_{t-i} + \varepsilon_t + \sum_{i=0}^{q} \lambda_i \varepsilon_{t-i} + \sum_{i=0}^{k} \gamma_i X_{t-i}$$

where the only difference between this equation and the above ARMAX model is X_t, referring to the composite index from the selected subset of search query data based on different feature selection methods.

4 In the last step, to evaluate whether our proposed forecasting models on the basis of four machine learning methods outperformed the benchmark ARMAX model, we compared the dynamic forecasting results of these different models by using the root-mean-square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and improvement ratio (IR) based on each measure. The evaluation measures were constructed using the following equations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y\_pred_i - y_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| y\_pred_i - y_i |$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}| y\_pred_i - y_i | / y_i$$

$$IR_{i,j} = 1 - \frac{RMSE_i (MAE_i / MAPE_i)}{RMSE_j (MAE_i / MAPE_i)}$$

where y_pred_i and y_i capture predicted tourist arrivals and actual tourist arrival data. $RMSE_{i/j}$, $MAE_{i/j}$, and $MAPE_{i/j}$ are the RMSEs, MAEs, and MAPEs of models *i* and *j*,

respectively. $IR_{i,j}$ measures how model *i* improves forecasting accuracy compared to model *j* in terms of the reduction of forecasting errors of RMSEs, MAEs, and MAPEs.

4. Empirical Results

We conducted an empirical study of forecasting monthly domestic tourist arrivals to Beijing, China, to examine the performance of the proposed methodology in selecting the most appropriate search query data and forecasting tourism demand.

First, Internet search query data from Baidu were collected relative to various aspects of tourism activities such as dining, lodging, traffic, recreation, shopping, and relevant tourist attractions. Second, four feature selection methods were applied to select the most appropriate subset of search query data. Third, we then constructed forecasting models using different types of selected search data as explanatory variables and compared the forecasting accuracy of the forecasting models with the benchmark model. Furthermore, we conducted one empirical experiment about the weekly forecasting of hotel occupancy in Charleston, SC to provide a robustness check about our methodology.

4.1 Data description

We select Beijing City as the target destination. Beijing as the capital of China has been developed as one famous tourism city, and the accurate forecasting of its tourist arrivals has attracted increasing attention from existing studies such as Li et al. (2017) and Sun et al. (2019).

Two kinds of data including monthly Beijing tourist arrivals and search query data were collected from January 2011 to August 2019, including 104 data points. Baidu search data were used since Baidu has the biggest market share in China compared to other search

engines (Yang et al. 2015). Our collected data sets can reflect the most recent trends in both tourist arrivals data and search query data. The tourist arrivals data were obtained from Wind database (http://www.wind.com.cn/), and the search query data were collected from Baidu; this search engine's data apparently perform better than Google's when forecasting domestic tourist arrivals in China (Yang et al. 2015).

The average number of tourist arrivals in Beijing (in 10,000) was 2,379 with standard deviations of 746.32. Figure 2 shows actual tourist arrival data for Beijing, which presents distinct cyclical characteristics.

[Please insert Figure 2 near here]

Baidu search query data were collected from Baidu's search engine on the basis of procedure of search keywords selection shown in Figure 1. Numerous keywords related to tourism activities were considered, which covered the aspects of tourists' dining, lodging, traffic, recreation, shopping, choices of attraction and other relevant activities. Several search query data were not collected given the low volume on Baidu search engine. Furthermore, famous attractions such as "The Great Wall" and "The Palace Museum" were included to reflect the tourists' attention on travelling in Beijing. In total, we obtained 59 search query data series. Figure 3 shows the related search keywords reflected in different categories.

[Please insert Figure 3 near here]

4.2 Selection of search query data

Four feature selection methods including filter-based feature selection, recursive feature selection, genetic algorithm feature selection, and random forest feature selection were used to select the appropriate subsets of search query data. All search query data were ranked, and less important data were eliminated.

We obtained four subsets of search query data for each method according to specific algorithms introduced in the Methodology section using R software. A common criterion when selecting subsets is that a feature will be eliminated if it fails to improve the forecasting accuracy. The detailed number of selected search query data was not necessarily the same given the discrepancy of the algorithms. Therefore, for the convenience of modelling and evaluation, only top five search query data series were selected to represent the information contained in the original data set for each feature selection method. Table 2 lists the selected search query data using four machine learning-based methods.

[Please insert Table 2 near here]

After obtaining the four groups of search query data, we computed the composite index using shift and summation process in Sun et al. (2019) to represent the linear combination of selected search query data. Therefore, the further analysis and modelling were conducted on the basis of the constructed indexes. Here, *Index0* was used to represent the index computed from all search query data without feature selection. In addition, *Index1*, *Index2*, *Index3*, and *Index4* represented the indexes obtained from four feature selection

methods. Figure 4 shows the tourist arrivals data and the five indexes. All data were standardized for the convenience of comparison in one figure.

[Please insert Figure 4 near here]

As shown in Figure 4, the graphic features between *Index0* and tourist arrivals data were quite different, while the other four data series were closely related to tourist arrivals data. To further analyze the relationships between the selected indexes and tourist arrivals data, we conducted the Pearson correlation analysis, Granger causality tests, and co-integration tests.

The Pearson correlation coefficients and relevant statistics were provided in Table 3. The correlation coefficient between the number of tourist arrivals and the index constructed from all search query data (*Index0*) without feature selection was the smallest among the five indexes. The correlation coefficient between *Index4* and the tourist arrivals data was 0.76. In particular, the indexes constructed from four feature selection methods have stronger correlation with the predicted tourist arrivals data. The result suggests that all feature selection methods can select more relevant search query data compared to that without feature selection.

[Please insert Table 3 near here]

4.3 Estimation results of forecasting models

To investigate whether and which machine-learning-based feature selection method could best improve the forecasting accuracy of tourism demand, the following forecasting models were constructed based on different subsets of search query data. First, a classical ARMA model without search query data or machine learning methods was constructed for tourism forecasting (Li et al. 2017; Pan, Wu, and Song 2012). We chose the ARMA model because we compared the performances among AR, ARMA, and ARIMA models, and found that an ARMA model can achieve the highest accuracy. Therefore, other models incorporating search query data were built upon the ARMA model.

Second, a benchmark ARMAX model incorporating a composite index of all search query data was constructed to explore whether machine-learning-based feature selection could outperform the model without feature selection. Third, we proposed four models using selected search datasets based on the following feature selection methods: filter-based feature selection, recursive feature selection, genetic algorithm feature selection, and random forest feature selection, which were noted as Models 1-4. We conducted a logarithmic transformation of tourist arrival data to reduce the impact of outliers. The lag orders of auto-regression term and moving-averaging term were specified on the basis of the AIC.

Tables 4-6 show the estimation results of the constructed six models. The differences among the models lie in the incorporation of the explanatory variables. For the ARMA model, only the AR and MA terms were considered to predict the tourist arrivals data. ARMAX model incorporated the *Index0* as the explanatory variables, which contained the

information from all search query data. However, Models 1-4 included the variables *Index1*, *Index2*, *Index3*, and *Index4* to reflect the extracted information from different feature selection methods.

[Please insert Table 4 near here]

[Please insert Table 5 near here]

[Please insert Table 6 near here]

As indicated from the above estimation results of six models, the constructed indexes are significant at 5% or 1% level. Different lag orders of the explanatory variables are selected according to AIC. In addition, Models 1-4 improves the adjusted R-squared compared to the ARMA and ARMAX model. Overall, search query data index built from feature selection methods could significantly predict the changes in tourist arrival data.

4.4 Comparisons of forecasting performance

To provide a robust and reliable evaluation about forecasting performance, Table 7 indicates one-step-ahead, two-step-ahead, and four-step ahead for Beijing tourism demand forecasting in terms of RMSE, MAE, MAPE, and IR. We particular focus on the improvement in forecasting accuracy when incorporating the search query data. First, the values of RMSE, MAE, and MAPE in the forecasting models showed that the ARMA model without search query data or feature selection as well as ARMAX model without feature selection methods exhibited relative larger forecasting errors than Models 1-4. For the

reduction of forecasting errors of RMSE, Models 1–4 with four feature selection methods significantly improved forecasting accuracy compared to ARMA model by 19.78%, 26.74%, 22.56%, and 21.87%, respectively in the one-step-ahead forecasting. Similar findings for two- and four-step ahead forecasts measured by MAE and MAPE can be achieved when compared to ARMA model.

Second, when compared to ARMAX model, the best forecasting model varies across different periods. For the one- and two-step ahead forecasting, Models 1–4 always outperformed the ARMAX model, indicated by the decreased forecasting errors in terms of RMSE, MAE, and MAPE. The average IRs (measured by RMSE) of the four models using different feature selection methods were 16.04% and 30.33%, for the one- and two-step ahead forecasting respectively. However, for the four-step-ahead forecasting measured by RMSE, Models 1, 3 and 4 still outperform the ARMAX model; while the ARMAX model is found to perform better than Model 2, suggesting that the feature selection methods do not necessarily improve the forecasting accuracy in the long run. The result further demonstrates that the proposed method is particularly effective for the short-term forecasting with search query data (Park, Lee, and Song 2017).

[Please insert Table 7 near here]

In general, the out-of-sample forecasting results indicated that models with feature selection methods significantly improved the forecasting accuracy of tourism demand in Beijing. To further answer the research question about which feature selection method can

improve forecasting performance to the greatest extent, we computed forecasting errors of RMSE, MAE, MAPE, and IR for each feature selection method.

Figure 5 depicts detailed forecasting errors from the four feature selection methods based on dynamic one-step-ahead out-of-sample forecasting. The overall performance of each method was not significantly different; the MAPEs of four models were 0.6295, 0.5410, 0.5939, and 0.5972, respectively. RMSEs in four models ranged from 0.0551 to 0.0604 with the mean of 0.0582 and standard deviation of 0.0022. MAEs ranged from 0.0415 to 0.0484 with the mean of 0.0453 and standard deviation of 0.0028. MAPEs ranged from 0.541 to 0.6295 with the mean of 0.5904 and standard deviation of 0.0366. Furthermore, the improvement ratios of four models ranged from 12.83 to 20.39. In summary, these findings suggest that the proposed feature selection methods can significantly improve forecasting accuracy, but we did not observe a significant difference in forecasting performance among these methods.

[Please insert Figure 5 near here]

4.5 Robustness with respect to hotel forecasting

The above empirical results have demonstrated the effectiveness of machine-learning-based feature selection in tourism demand forecasting. We also examined whether machine-learning-based feature selection methods could improve forecasting in a specific dimension of tourism demand: hotel occupancy. We chose Charleston, SC as the target destination because the hotel occupancy data are accessible, which were gathered weekly from January

2006 to February 2014 including 426 data points (Pan and Yang 2017). Here, we briefly discuss the data description, feature selection results, estimation results of forecasting models, and forecasting evaluation.

Consistent with Pan and Yang (2017), search query data were gathered from Google, including 45 search keywords related to hotel occupancy in Charleston. Figure 6 illustrates weekly hotel occupancy data in the city; the average hotel occupancy rate was 0.7028. These data demonstrated significant cyclical characteristics.

[Please insert Figure 6 near here]

Following the proposed methodology, we selected weekly Google search data using the four feature selection methods and incorporated these data into forecasting models: one benchmark ARMAX model and four machine-learning-based models. The forecasting models were constructed for weekly forecasting of hotel occupancy in Charleston. To be consistent with Pan and Yang (2017), we constructed the benchmark ARMAX model by including the search query data taking the keyword "hotel Charleston" as the explanatory variable. For Hotel-Models 1–4, we incorporated the aggregated index from chosen feature selection methods as the explanatory variable, consistent with our practice in the aforementioned Beijing forecasting study.

Table 8 displays the estimation results of five econometric models on forecasting weekly hotel occupancy. The dependent variable was the weekly hotel occupancy rate in Charleston. The lag orders of autoregressive and moving average terms were decided based on AIC. The explanatory variables including AR(1), MA(4), and the selected search query

data are significant at the 1% significance level. All four forecasting models with feature selection methods showed an increase in the adjusted R-squared.

[Please insert Table 8 near here]

Table 9 shows the one-, two-, and four-step-ahead hotel occupancy forecasting evaluation in terms of RMSE, MAE, MAPE and IR values. Forecasting models with feature selection methods exhibited lower values of RMSE, MAE, and MAPE than the benchmark model. Compared to the benchmark ARMAX model, the four forecasting models reduced the RMSEs by 18.4%, 16.41%, 17.04%, and 20.13%, respectively in one-step-ahead forecasting; the average improvement in forecasting accuracy was 18.00%. These results suggest that machine learning methods can select useful subsets of search query data to significantly improve hotel occupancy forecasting. Figure 7 depicts the improvement of forecasting accuracy of four weekly forecasting models compared to the benchmark ARMAX model. The findings were in line with those obtained in the monthly tourism forecasting study, which suggested that the proposed four feature selection methods could improve the forecasting accuracy with insignificant differences.

[Please insert Table 9 near here]

[Please insert Figure 7 near here]

5. Conclusions

Accurate and timely forecasting presents a crucial challenge for industries and academia in tourism and hospitality (Song, Qiu, and Park 2019). This study investigated whether machine-learning-based methods can select the most useful combinations of search keywords to improve tourism forecasting accuracy. We applied four machine-learning-based feature selection methods (i.e., filter-based feature selection, recursive feature selection, genetic algorithm feature selection, and random forest feature selection). Forecasting performance of the proposed methods was calculated based on tourism demand in Beijing as well as hotel occupancy rates in Charleston, SC.

Our findings indicate that a useful subset of search query data can be obtained via machine-learning-based methods, which can in turn significantly improve forecasting accuracy. The empirical studies based on forecasting monthly tourist arrivals and weekly hotel occupancy indicate superior performance of the proposed feature selection methods. The results of our forecasting evaluation did not reveal a significant difference among the four feature selection methods in terms of reduced RMSE, MAE, and MAPE. We found that these methods could obtain optimal subsets of search query data to improve forecasting accuracy.

Our study contributes to the relevant literature by implementing feature selection to balance the coverage of search query data and forecasting accuracy. To the best of our knowledge, this study is the first to apply different feature selection methods to extract useful

information from Internet search query data. Selected subsets of search query data from all feature selection methods outperformed the ARMAX model without feature selection. These findings further confirm that the selected keywords related to dining, lodging, traffic, recreation, shopping, tourism and attraction are effective in tourism forecasting. Furthermore, the value of search query data increases with the adoption of machine-learning-based feature selection methods. In particular, one significant advantage of the proposed feature selection methods is the automatic selection of useful search query data for effectively predicting tourism demand. The overfitting problems that are usually caused by a large data set can be avoided, which contributes to the improvement of forecasting accuracy. Such a 'bottom-up' strategy of feature selection methods used for search query data entails selecting keywords automatically and less depending on intuition or prior knowledge.

This study has several limitations. First, although we explored the performance of four feature selection methods in selecting search query data, other methods (e.g., least absolute shrinkage and selection operator and principle component analysis) could be used to select and consolidate large volumes of search query data based on a linear regression framework (Song and Liu 2017; Song, Qiu, and Park 2019). Therefore, future studies could compare these approaches with our proposed machine-learning-based methods in selecting search query data. Second, due to the data availability issue, the forecasting of tourist arrivals from different source markets was not conducted in this study, which could be addressed in future studies to provide more support for destination management. Furthermore, this study only considered search query data for tourism and hospitality forecasting. Other big data,

such as from social media, should be considered to improve forecasting accuracy. Therefore, subsequent research could extend feature selection methods by integrating more user-generated big data from multiple Internet platforms to improve tourism and hospitality forecasting accuracy.

References

Assaf, A. G., G. Li, H. Song, and M. G. Tsionas. 2019. "Modeling and Forecasting Regional Tourism Demand Using the Bayesian Global Vector Autoregressive (BGVAR) Model." *Journal of Travel Research* 58 (3): 383-397.

Bangwayo-Skeete, P. F., and R. W. Skeete. 2015. "Can Google Data Improve the Forecasting Performance of Tourist Arrivals? Mixed-data Sampling Approach." *Tourism Management, 46*: 454-464.

Bokelmann, B., and S. Lessmann. 2019. "Spurious Patterns in Google Trends Data - An Analysis of the Effects on Tourism Demand Forecasting in Germany." *Tourism Management* 75: 1–12.

Breiman, L. 2001. "Random forests." *Machine Learning* 45(1): 5-32.

Camacho, M., and M. J. Pacce. 2018. "Forecasting Travellers in Spain with Google' s Searches." *Tourism Economics* 24(4): 434-448.

Brynjolfsson, E., T. Geva, and S. Reichman. 2016. "Crowd-squared: Amplifying the predictive power of search trend data." *MIS Quarterly* 40(4) : 941-961.

Cai, J., J. Luo. S. Wang, and S. Yang. 2018. "Feature Selection in Machine Learning: A New Perspective." *Neurocomputing* 300: 70-79.

Chandrashekar, G., and F. Sahin. 2014. "A Survey on Feature Selection Methods." *Computers & Electrical Engineering* 40(1): 16-28.

Chen, J. L., G. Li, D. C. Wu, and S. Shen. (2019). "Forecasting Seasonal Tourism Demand Using a Multiseries Structural Time Series Method." *Journal of Travel Research* 58(1): 92-103.

Choi, H., and H. Varian. 2012. "Predicting the Present with Google Trends." *Economic Record* 88(s1): 2-9.

Cui, R., S. Gallion, A. Moreno, and D. Zhang. 2017. "Operational Value of Social Media Information." *Production and Operations Management* 27(10): 1749–1769.

Das, A. K., S. Das, and A. Ghosh. 2017. "Ensemble Feature Selection using Bi-objective Genetic Algorithm." *Knowledge-Based Systems* 123: 116-127.

Domingos, P. 2012. "A Few Useful Things to Know About Machine Learning." *Communications of the ACM* 55(10): 78-87.

Fesenmaier, D. R., Z. Xiang, B. Pan, and R. Law. 2011. "A Framework of Search Engine Use for Travel Planning." *Journal of Travel Research* 50 (6): 587–601.

Frechtling, D. 2012. *Forecasting Tourism Demand.* New York: Routledge.

Genuer, R., J. M. Poggi, and C. Tuleau-Malot. 2010. "Variable Selection Using Random Forests." *Pattern Recognition Letters* 31(14): 2225-2236.

Geva, T., G. Oestreicher-Singer, N. Efron, and Y. Shimshoni. 2017. "Using Forum and Search Data For Sales Prediction of High-involvement Products." *MIS Quarterly* 41(1): 65-82.

Guizzardi, A., and A. Stacchini. 2015. "Real-time Forecasting Regional Tourism with Business Sentiment Surveys." *Tourism Management* 47: 213-223.

Gunter, U. and I. Önder. 2015. “Forecasting International City Tourism Demand for Paris: Accuracy of Uni- and Multivariate Models Employing Monthly Data.” *Tourism Management* 46: 123-135.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. “Gene Selection For Cancer Classification Using Support Vector Machines.” *Machine Learning* 46(1-3): 389-422.

Guyon, I., and A. Elisseeff. 2003. “An Introduction To Variable and Feature Selection.” *Journal of Machine Learning Research* 3(Mar): 1157-1182.

Hassani, H., E. S. Silva, N. Antonakakis, G. Filis, and R. Gupta. 2017. “Forecasting Accuracy Evaluation of Tourist Arrivals.” *Annals of Tourism Research* 63: 112–127.

Huang, C. L., and C. J. Wang. 2006. “A GA-based Feature Selection and Parameters Optimizationfor Support Vector Machines.” *Expert Systems with applications* 31(2): 231-240.

Huang, X., L. Zhang., and Y. Ding. 2017. “The Baidu Index: Uses in Predicting Tourism Flows –A Case Study of the Forbidden City.” *Tourism Management* 58: 301–306.

Hu, M., and H. Song. 2019. “Data Source Combination for Tourism Demand Forecasting.” *Tourism Economics*. doi: 10.1177/1354816619872592.

Kodratoff, Y., and R. S. Michalski. 2014. *Machine Learning: An Artificial Intelligence Approach (Vol. 3).* San Francisco: Morgan Kaufmann.

Law, R., G. Li, D. K. C. Fong, and X. Han. 2019. “Tourism Demand Forecasting: A Deep Learning Approach.” *Annals of Tourism Research* 75: 410-423.

Li, J., L. Xu. L, Tang, S. Wang, and L. Li. 2018a. “Big Data in Tourism Research: A Literature Review.” *Tourism Management* 68: 301-323.

Li, S., T. Chen, L. Wang, and C. Ming. 2018b. “Effective Tourist Volume Forecasting Supported by PCA and Improved BPNN Using Baidu Index.” *Tourism Management* 68: 116–126.

Li, X., and Law, R. 2020. “Forecasting Tourism Demand with Decomposed Search Cycles.” *Journal of Travel Research* 59 (1): 52-68.

Li, X., B. Pan, R. Law, and X. Huang. 2017. “Forecasting Tourism Demand with Composite Search Index.” *Tourism Management* 59: 57-66.

Long, W., C. Liu, and H. Song. 2019. “Pooling in Tourism Demand Forecasting.” *Journal of Travel Research* 58(7): 1161-1174.

Kursa, M. B., and W. R. Rudnicki. 2010. “Feature Selection with the Boruta Package.” *Journal of Statistical Software* 36(11): 1-13.

Önder, I. 2017. “Forecasting Tourism Demand with Google Trends: Accuracy Comparison of Countries versus Cities.” *International Journal of Tourism Research* 19 (6): 648–660.

Pai, P. F., K. C. Hung, and K. P. Lin. 2014. “Tourism Demand Forecasting Using Novel Hybrid System.” *Expert Systems with Applications* 41(8): 3691-3702.

Pan, B., D. C. Wu, and H. Song. 2012. “Forecasting Hotel Room Demand Using Search Engine Data.” *Journal of Hospitality and Tourism Technology* 3(3): 196-210.

Pan, B., and Y. Yang. 2017. “Forecasting Destination Weekly Hotel Occupancy With Big Data.” *Journal of Travel Research* 56(7): 957-970.

Peng, B., H. Song. and G. I. Crouch. 2014. “A Meta-analysis of International Tourism Demand Forecasting and Implications for Practice.” *Tourism Management* 45: 181-193.

Rivera, R. 2016. “A Dynamic Linear Model to Forecast Hotel Registrations in Puerto Rico Using Google Trends Data.” *Tourism Management* 57: 12-20.

Saeys, Y., I. Inza, and P. Larrañaga. 2007. “A Review of Feature Selection Techniques in Bioinformatics.” *Bioinformatics* 23(19): 2507-2517.

Shen, S., G. Li, and H. Song. 2011. “Combination Forecasts of International Tourism Demand.” *Annals of Tourism Research* 38(1): 72–89.

Song, H., and G. Li. 2008. “Tourism Demand Modelling and Forecasting-A Review of Recent Research.” *Tourism Management* 29(2): 203–220.

Song, H., and H. Liu. 2017. “Predicting Tourist Demand using Big Data.” In Z. Xiang & D. R. Fesenmaier (Eds.), *Analytics in Smart Tourism Design: Concepts and Methods* (13-29), Springer, Cham.

Song, H., R. T. Qiu, and J. Park. 2019. “A Review of Research on Tourism Demand Forecasting.” *Annals of Tourism Research* 75: 338-362.

Song, H., and S. F. Witt. 2000. *Tourism Demand Modeling and Forecasting: Modern Econometric Approaches*. Oxford: Pergamon Press.

Sun, X., W. Sun, J. Wang, Y. Zhang, and Y. Gao. 2016. “Using A Grey–Markov Model Optimized by Cuckoo Search Algorithm to Forecast the Annual Foreign Tourist Arrivals to China.” *Tourism Management* 52: 369-379.

Sun, S., Y. Wei, K. L. Tsui, and S. Wang. 2019. “Forecasting Tourist Arrivals with Machine Learning and Internet Search Index.” *Tourism Management* 70: 1–10.

Sylvester, E. V., P. Bentzen, I. R. Bradbury, M. Clément, et al. 2018. “Applications of Random Forest Feature Selection for Fine‐scale Genetic Population Assignment.” *Evolutionary Applications* 11(2): 153-165.

Valadkhani, A., and B. O'Mahony. 2018. “Identifying Structural Changes and Regime Switching in Growing and Declining Inbound Tourism Markets in Australia.” *Current Issues in Tourism* 21(3): 277-300.

Wen, L., C. Liu., and H. Song. 2019. “Forecasting Tourism Demand Using Search Query Data: A Hybrid Modelling Approach.” *Tourism Economics* 25(3): 309-329.

Xie, G., X. Li, Y. Qian, and S. Wang. 2020. “Forecasting Tourism Demand with KPCA-based Web Search Indexes.” *Tourism Economics*. doi: 10.1177/1354816619898576.

Xue, X., M. Yao, and Z. Wu. 2018. “A Novel Ensemble-based Wrapper Method for Feature Selection Using Extreme Learning Machine and Genetic Algorithm.” K*nowledge and Information Systems* 57(2): 389-412.

Yan, K., and D. Zhang. 2015. “Feature Selection and Analysis on Correlated Gas Sensor Data with Recursive Feature Elimination.” *Sensors and Actuators B: Chemical* 212: 353-363.

Yang, Y., and H. Zhang. 2019. “Spatial-temporal Forecasting of Tourism Demand.” *Annals of Tourism Research* 75: 106-119.

Yang, X., B. Pan, J. A. Evans, and B. Lv. 2015. “Forecasting Chinese Tourist Volume with Search Engine Data.” *Tourism Management* 46: 386-397.

Yu, L., and H. Liu. 2004. “Efficient Feature Selection via Analysis of Relevance and Redundancy.” *Journal of Machine Learning Research* 5(Oct): 1205-1224.

Zhang, B., X. Huang., N. Li., and R. Law. 2017. “A Novel Hybrid Model for Tourist Volume Forecasting Incorporating Search Engine Data.” *Asia Pacific Journal of Tourism Research* 22(3): 245-254.

Zhou-Grundy, Y., and L. W. Turner. 2014. “The Challenge of Regional Tourism Demand Forecasting: The Case of China.” *Journal of Travel Research* 53(6): 747-759.

Appendix A. Tables

Table A.1 Recursive feature selection algorithm

1. Train the linear model using all search query data
2. Calculate the model performance such as RMSE
3. Calculate the variable rankings
4. **for** Each subset of search data S_i, do

   Keep the S_i most important variables

   Train the model on the training set using S_i variables

   Calculate model performance

   end
5. Calculate the performance over the S_i
6. Decide the appropriate number of search data

Table A.2 Random forest selection procedure

1. Compute the OOB error_0 for each tress in the random forest;

2. Randomly select one feature and add noise to it, compute the OOB error_1;

3. The importance of the selected feature is computed using:

$$\sum(OOBerror\_1 - OOBerror\_0).$$

The importance of the feature is obtained using the difference between the two errors. If the OOB error_1 largely reduces by adding the noise, the feature is important for the overall performance.

4. Rank the features based on the computed importance;

5. Repeat the above steps until a subset of features are selected.

Table 1. Overview of selected papers with search query data

Paper	Search engines	Search query data	Forecasting models	Predicted context
Bangwayo-Skeete and Skeete (2015)	Google	Search query index for "hotels and flights"	AR-MIDAS	Five destinations in the Caribbean
Bokelmann and Lessmann (2019)	Google	269 search terms related to a holiday	Seasonal ARIMA	Tourist arrivals in several German holiday regions
Camacho and Pacce (2018)	Google	Search query indices for travelling related topics	DFA	Overnight stays of travelers in Spain
Hu and Song (2019)	Google	Four aspects related to Macau tourism: dining, lodging, transportation, and recreation	BPNN	Short-haul travel from Hong Kong to Macau
Huang, Zhang and Ding (2017)	Baidu	Five categories of destination related terms	ARIMA, ADL	Tourist arrivals to the Forbidden city, Beijing
Law et al. (2019)	Google	Tourism related keywords from seven categories	DL	Monthly Macau tourist arrivals
Li et al. (2018b)	Baidu	15 relevant terms with destination	PCA-ADE-BPNN	Tourist arrivals to Beijing and Hainan Province, China
Li and Law (2020)	Google	Search query index for "Hong Kong" under Vacation Destinations	EEMD-ARX	Tourist arrivals to Hong Kong
Li et al. (2017)	Baidu	46 relevant terms with Beijing tourism	GDFM-ARX	Tourist arrivals to Beijing China
Önder (2017)	Google	Search query indices for Austria, Belgium, Vienna, and Barcelona	ADL	Tourist arrivals to Vienna, Barcelona, Austria and Belgium
Pan and Yang (2017)	Google	A relevant search term: Charleston hotels	ARMAX	Hotel occupancy rate of Charleston
Pan, Wu, and Song (2012)	Google	Five relevant terms: Charleston SC, travel Charleston, Charleston hotels, Charleston restaurants, and Charleston tourism	ARIMA, ADL, TVP, VAR	Hotel room demand in Charleston, SC
Park, Lee, and Song (2017)	Google	31 terms related with South Korea tourism	Seasonal ARIMA	Tourist inflow of Japanese tourists to South Korea
Rivera (2016)	Google	Search query indices for travel search categories	DLM	Hotel nonresident registrations in Puerto Rico
Wen, Liu, and Song (2019)	Baidu	64 relevant keywords with Hong Kong tourism	Hybrid models of ARIMAX and NARX	Tourist arrivals in Hong Kong from mainland China

Sun et al. (2019)	Google, Baidu	22 relevant keywords with Beijing tourism	KELM	Tourist arrivals to Beijing
Xie et al. (2020)	Google, Baidu	78 relevant keywords with Hong Kong tourism	KPCA	Tourist arrivals to Hong Kong
Zhang et al. (2017)	Baidu	11 relevant keywords with Hainan tourism	Hybrid models of BA and SVR	Tourist arrivals to Hainan Province, China

Note. ARIMA: Autoregressive integrated moving average; AR-MIDAS: Autoregressive-mixed data sampling; ADE: Adaptive differential evolution; ADL: Autoregressive distributed lag; BA: Bat algorithm; BPNN: Back-propagation neural networks; DFA: Dynamic factor approach; DL: Deep learning; DLM: Dynamic linear model; EEMD: Ensemble empirical mode decomposition; GDFM: Generalized dynamic factor model; KELM: Kernel extreme learning machine; KPCA: Kernel principle component analysis; SVR: Support vector regression; VAR: Vector autoregressive.

Table 2. Selected Beijing search data with four feature selection methods

Feature selection method	Top 5 selected search query data
Method 1 Filter-based selection	Food Lodging Vocational village Rural tourism Hotel reservation
Method 2 Recursive feature selection	Food Travel strategy Fun places in Beijing Trip Ming tombs
Method 3 Genetic algorithm feature selection	Special food Hotel reservation Hotel prices Tourism Rong Bao Zhai
Method 4 Random forest feature selection	Food Rural tourism Hotel prices Travel strategy Tourism attraction

Table 3. Correlation analysis between tourist arrivals and five indexes

Tourist *vs*	Index0	Index1	Index2	Index3	Index4
	0.3831	0.5530	0.7398	0.5844	0.7619
t-Statistic	4.1892	6.7039	11.1047	7.2725	11.8804
Prob.	0.0001***	0.0000***	0.0000***	0.0000***	0.0000***

Note. *** indicates the significance level at 1%.

Table 4. Estimation results of ARMA and ARMAX models

ARMA model				
Dependent variable	Monthly tourist arrivals to Beijing (in log)			
Machine learning method	None feature selection methods			
Search query data	None search query data			
	Coefficient	Std. Error	t-Statistic	Prob.
Constant	8.4027	0.1735	48.4288	0.0000***
AR(2)	-0.0394	0.0120	-3.2941	0.0014***
AR(12)	0.9825	0.0120	82.1539	0.0000***
MA(12)	-0.8950	0.0330	-27.0958	0.0000***
Statistics				
R-squared	0.9474	Akaike info criterion		-2.1915
Adjusted R-squared	0.9456	Schwarz criterion		-2.0819
ARMAX model				
Dependent variable	Monthly tourist arrivals to Beijing (in log)			
Machine learning method	None feature selection methods			
Search query data	Index0			
	Coefficient	Std. Error	t-Statistic	Prob.
C	8.4660	0.1740	48.6540	0.0000***
Index0	0.0001	0.0000	-1.7149	0.0900*
Index0(-1)	0.0001	0.0000	2.0148	0.0471**
AR(12)	0.9739	0.0140	69.5492	0.0000***
AR(1)	-0.0283	0.0138	-2.0439	0.0441**
MA(12)	-0.9153	0.0191	-47.8672	0.0000***
Statistics				
R-squared	0.9556	Akaike info criterion		-2.3259
Adjusted R-squared	0.9529	Schwarz criterion		-2.1604

Note. *, **, *** indicate significance at the 10%, 5%, and 1% level, respectively.

Table 5. Estimation results of Models 1-2

Model 1: tourism forecasting combined with filter-based selection				
Dependent variable	Monthly tourist arrivals to Beijing (in log)			
Machine learning method	Filter-based selection			
Search query data	Index1			
	Coefficient	Std. Error	t-Statistic	Prob.
Index1	0.0003	0.0001	2.0853	0.0402**
Index1(-4)	-0.0004	0.0001	-2.9158	0.0046***
Constant	8.5603	0.2296	37.2838	0.0000***
AR(2)	-0.0137	0.0117	-1.1670	0.2466
AR(12)	0.9668	0.0109	88.4366	0.0000***
MA(12)	-0.9199	0.0179	-51.5334	0.0000***
Statistics				
R-squared	0.9660	Akaike info criterion		-2.6272
Adjusted R-squared	0.9639	Schwarz criterion		-2.4583
Model 2: tourism forecasting combined with recursive feature selection				
Dependent variable	Monthly tourist arrivals to Beijing (in log)			
Machine learning method	Recursive feature selection			
Search query data	Index2			
	Coefficient	Std. Error	t-Statistic	Prob.
Index2(-7)	0.0001	0.0000	2.7426	0.0076***
Index2(-12)	0.0000	0.0000	-0.6991	0.4867
Constant	8.3409	0.2051	40.6691	0.0000***
AR(2)	-0.0236	0.0096	-2.4574	0.0163**
AR(12)	0.9813	0.0094	104.5655	0.0000***
MA(12)	-0.9388	0.0171	-54.9246	0.0000***
Statistics				
R-squared	0.9705	Akaike info criterion		-2.7724
Adjusted R-squared	0.9686	Schwarz criterion		-2.5938

Note. **, *** indicate significance at the 5% and 1% level, respectively.

Table 6. Estimation results of Models 3-4

Model 3: tourism forecasting combined with genetic algorithm feature selection				
Dependent variable	Monthly tourist arrivals to Beijing (in log)			
Machine learning method	Genetic algorithm feature selection			
Search query data	Index3			
	Coefficient	Std. Error	t-Statistic	Prob.
Index3(-8)	0.0003	0.0001	3.7681	0.0003***
Constant	8.3221	0.1671	49.8070	0.0000***
AR(2)	-0.0171	0.0096	-1.7896	0.0773*
AR(12)	0.9602	0.0093	103.5438	0.0000***
MA(12)	-0.9340	0.0179	-52.2113	0.0000***
Statistics				
R-squared	0.9694	Akaike info criterion		-2.7128
Adjusted R-squared	0.9679	Schwarz criterion		-2.5681
Model 4: tourism forecasting combined with random forest feature selection				
Dependent variable	Monthly tourist arrivals to Beijing (in log)			
Machine learning method	Random forest feature selection			
Search query data	Index4			
	Coefficient	Std. Error	t-Statistic	Prob.
Index4(-8)	0.0002	0.0001	2.7068	0.0083**
Index4(-1)	-0.0001	0.0000	-2.1741	0.0327**
Constant	8.5204	0.2379	35.8129	0.0000***
AR(2)	-0.0122	0.0092	-1.3276	0.1882
AR(12)	0.9745	0.0092	106.4189	0.0000***
MA(12)	-0.9352	0.0175	-53.3526	0.0000***
Statistics				
R-squared	0.9687	Akaike info criterion		-2.6651
Adjusted R-squared	0.9667	Schwarz criterion		-2.4915

Note. *, **, *** indicate significance at the 10%, 5%, and 1% level, respectively.

Table 7. Evaluation of Beijing tourism demand forecasting

Evaluation criteria		**Model**					
Forecasting step		**ARMA**	**ARMAX**	**Model 1**	**Model 2**	**Model 3**	**Model 4**
one-step	RMSE	0.0752	0.0692	0.0604	0.0551	0.0583	0.0588
	MAE	0.0538	0.0529	0.0484	0.0415	0.0455	0.0458
	MAPE	0.7068	0.6943	0.6295	0.5410	0.5939	0.5972
two-step	RMSE	0.1136	0.1120	0.0675	0.0781	0.0932	0.0733
	MAE	0.0986	0.0955	0.0541	0.0571	0.0673	0.0571
	MAPE	1.3517	1.3115	0.7456	0.7924	0.9334	0.7889
four-step	RMSE	0.1119	0.0842	0.0619	0.0928	0.0779	0.0784
	MAE	0.1039	0.0679	0.0548	0.0792	0.0626	0.0702
	MAPE	1.3744	0.9124	0.7276	1.0463	0.8373	0.9247
IR: Five models vs ARMA model							
			ARMAX	**Model 1**	**Model 2**	**Model 3**	**Model 4**
one-step	IR_RMSE (%)		7.98	19.78	26.74	22.56	21.87
	IR_MAE (%)		1.70	10.18	22.91	15.41	14.95
	IR_MAPE (%)		1.77	10.94	23.46	15.98	15.51
two-step	IR_RMSE (%)		1.40	40.58	31.22	17.93	35.48
	IR_MAE (%)		3.09	45.14	42.04	31.78	42.06
	IR_MAPE (%)		2.97	44.84	41.37	30.94	41.64
four-step	IR_RMSE (%)		24.79	44.64	17.08	30.35	29.93
	IR_MAE (%)		34.63	47.27	23.77	39.79	32.47
	IR_MAPE (%)		33.62	47.06	23.87	39.08	32.72
IR: Modes 1-4 vs ARMAX model							
				Model 1	**Model 2**	**Model 3**	**Model 4**
one-step	IR_RMSE (%)			12.83	20.39	15.85	15.10
	IR_MAE (%)			8.62	21.57	13.95	13.47
	IR_MAPE (%)			9.34	22.09	14.47	13.99
two-step	IR_RMSE (%)			39.73	30.24	16.76	34.57
	IR_MAE (%)			43.39	40.19	29.61	40.21
	IR_MAPE (%)			43.15	39.58	28.83	39.85
four-step	IR_RMSE (%)			26.40	-10.24	7.40	6.84
	IR_MAE (%)			19.33	-16.62	7.89	-3.31
	IR_MAPE (%)			20.26	-14.68	8.23	-1.35

Table 8. Estimation results for weekly Charleston occupancy forecasting

ARMAX Model:				**Hotel-Model 1:**			
None feature selection				**Filter-based selection**			
	Coefficient	Std. Error	Prob.		Coefficient	Std. Error	Prob.
Constant	0.6926	0.0120	0.0000***	Constant	0.6949	0.0087	0.0000***
AR(1)	0.6428	0.0399	0.0000***	AR(1)	0.5451	0.0439	0.0000***
MA(4)	0.3767	0.0478	0.0000***	MA(4)	0.3425	0.0482	0.0000***
Search data	0.0429	0.0057	0.0000***	Search data	0.0750	0.0090	0.0000***
Search data(-4)	0.0259	0.0058	0.0000***	Search data(-4)	0.0545	0.0088	0.0000***
Adjusted R-Squared			0.7483	Adjusted R-Squared			0.7766
Akaike info criterion			-2.6447	Akaike info criterion			-2.7636
Hotel-Model 2:				**Hotel-Model 3:**			
Recursive feature selection				**Genetic algorithm feature selection**			
	Coefficient	Std. Error	Prob.		Coefficient	Std. Error	Prob.
Constant	0.6848	0.0093	0.0000***	Constant	0.6899	0.0089	0.0000***
AR(1)	0.5817	0.0437	0.0000***	AR(1)	0.5478	0.0435	0.0000***
MA(4)	0.3285	0.0483	0.0000***	MA(4)	0.3423	0.0481	0.0000***
Search data	0.0876	0.0096	0.0000***	Search data	0.0636	0.0090	0.0000***
Search data(-4)	0.0488	0.0092	0.0000***	Search data(-4)	0.0606	0.0088	0.0000***
Adjusted R-Squared			0.7832	Adjusted R-Squared			0.7685
Akaike info criterion			-2.7937	Akaike info criterion			-2.7281
Hotel-Model 4:							
Random forest feature selection							
	Coefficient	Std. Error	Prob.				
Constant	0.6925	0.0083	0.0000***				

AR(1)	0.5194	0.0441	0.0000***
MA(4)	0.3348	0.0481	0.0000***
Search data	0.0627	0.0088	0.0000***
Search data(-4)	0.0646	0.0087	0.0000***
Adjusted R-Squared		0.7675	
Akaike info criterion		-2.7241	

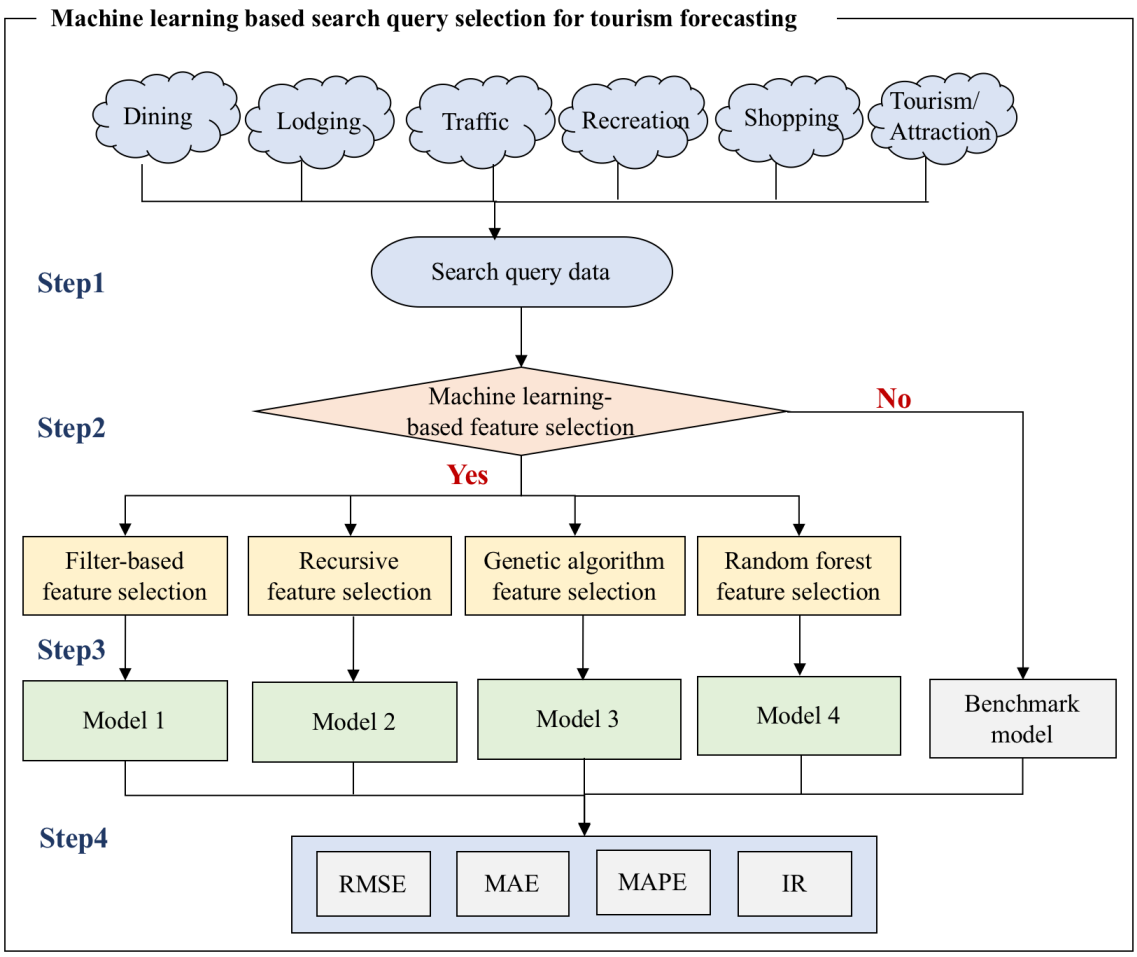
Note. ******* indicates significance at the 1% level.

Table 9. Evaluation of hotel occupancy forecasting models

Evaluation criteria		**Model**				
Forecasting step		**ARMAX**	**Hotel-Model 1**	**Hotel-Model 2**	**Hotel-Model 3**	**Hotel-Model 4**
one-step	RMSE	0.0966	0.0788	0.0808	0.0802	0.0772
	MAE	0.0786	0.0649	0.0664	0.0646	0.0623
	MAPE	12.8020	10.1450	10.3443	10.2253	9.8022
two-step	RMSE	0.1062	0.0907	0.0884	0.1033	0.0796
	MAE	0.1011	0.0847	0.0794	0.0967	0.0793
	MAPE	15.4535	12.9359	12.1095	14.7621	13.9218
four-step	RMSE	0.0976	0.0518	0.0581	0.0517	0.0480
	MAE	0.0821	0.0414	0.0532	0.0445	0.0402
	MAPE	16.8534	8.3457	10.4697	8.8748	8.1948

IR: Hotel-Models 1-4 vs ARMAX model

		Hotel-Model 1	**Hotel-Model 2**	**Hotel-Model 3**	**Hotel-Model 4**
one-step	IR_RMSE (%)	18.40	16.41	17.04	20.13
	IR_MAE (%)	17.44	15.59	17.86	20.75
	IR_MAPE (%)	20.75	19.20	20.13	23.43
two-step	IR_RMSE (%)	14.59	16.70	2.68	25.00
	IR_MAE (%)	16.25	21.54	4.43	21.57
	IR_MAPE (%)	16.29	21.64	4.47	9.91
four-step	IR_RMSE (%)	46.87	40.50	47.03	50.87
	IR_MAE (%)	49.51	35.13	45.84	51.00
	IR_MAPE (%)	50.48	37.88	47.34	51.38



Note. Step1-search keywords selection; Step2-machine learning based feature selection; Step3-econometric modelling; Step4-forecasting evaluation

Figure 1. Proposed forecasting framework

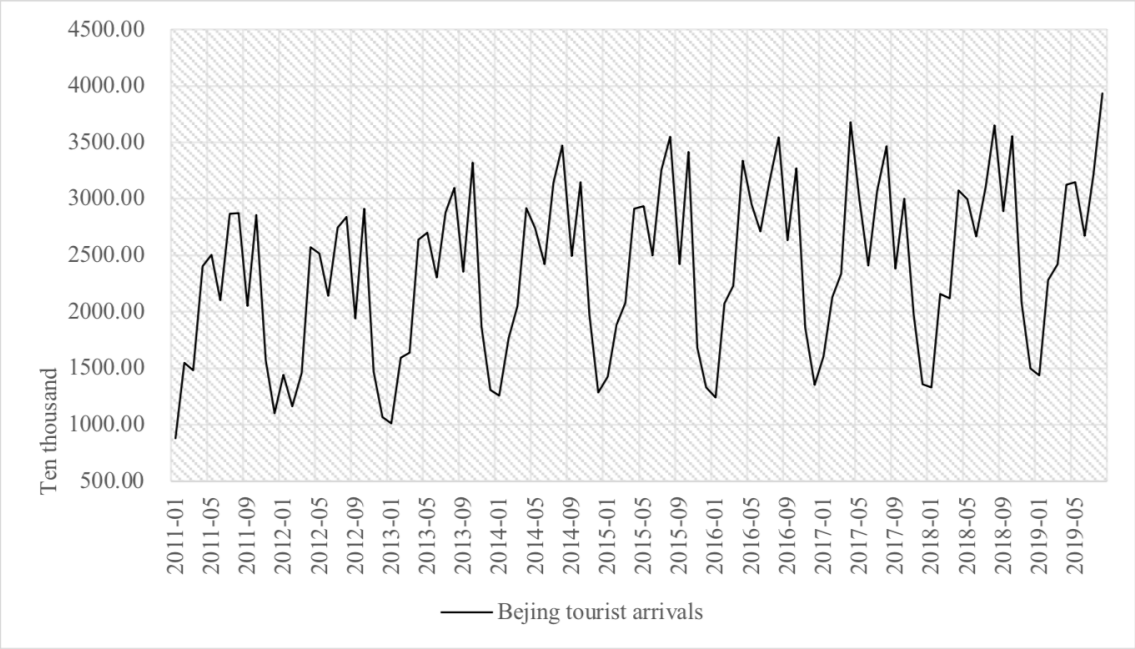


Figure 2. Monthly tourist arrivals in Beijing (2011.1-2019.8)

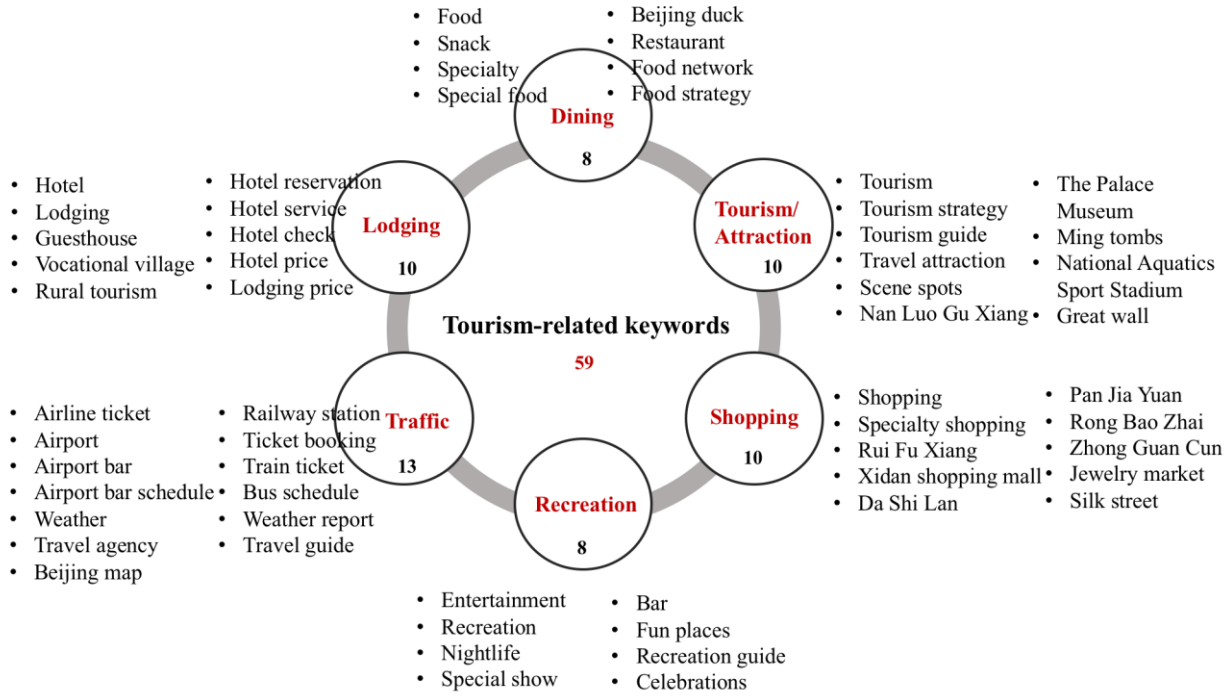


Figure 3. Baidu search keywords for Beijing tourism

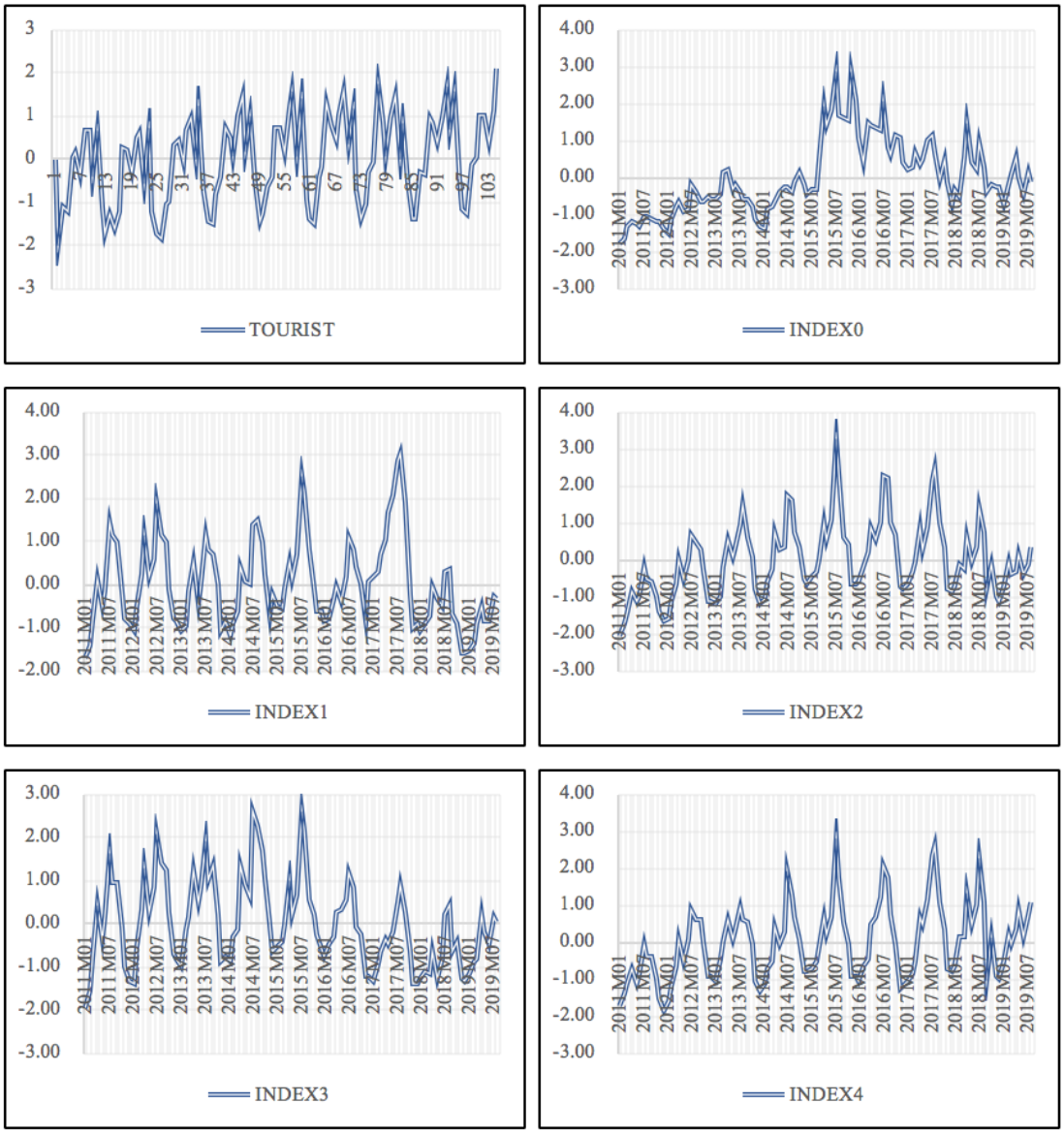


Figure 4. Tourist arrivals data and the five indexes

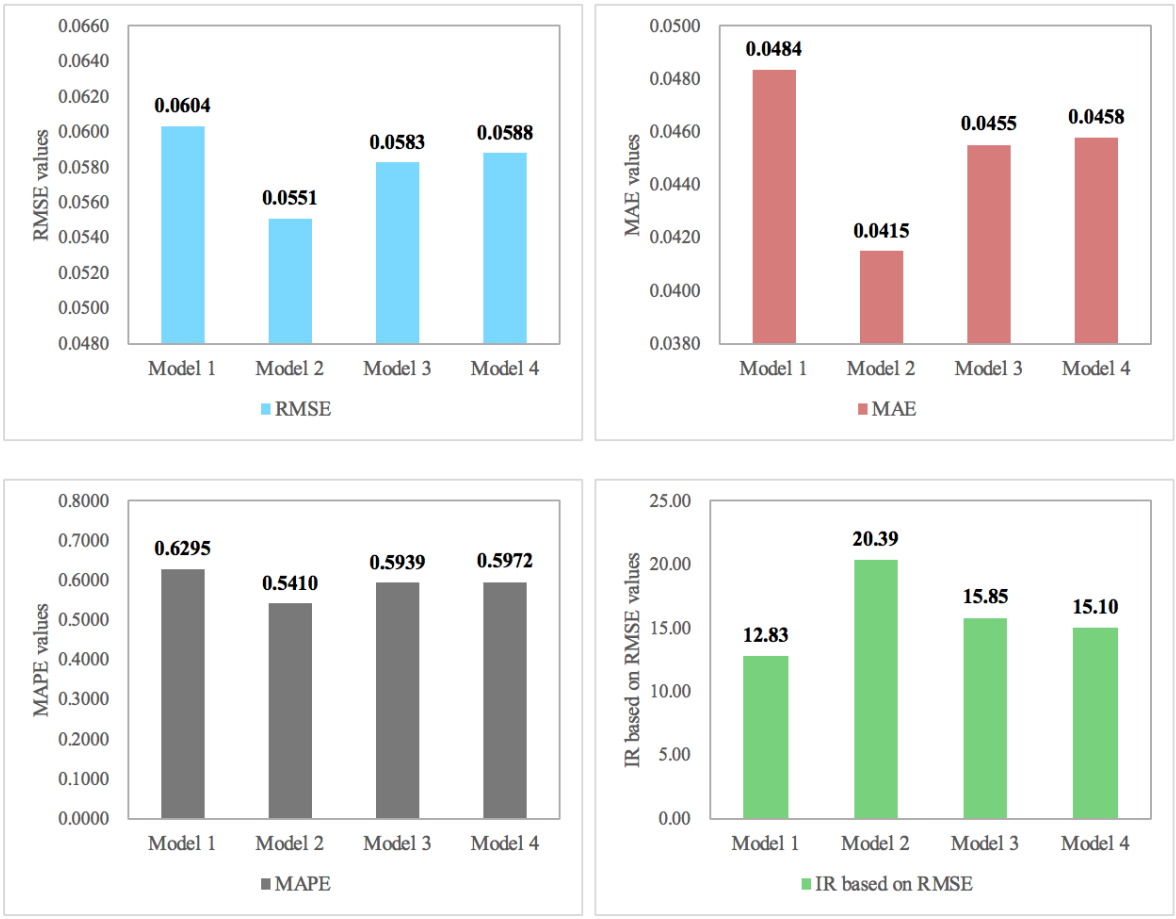


Figure 5. RMSE, MAE, MAPE, and IR among Models 1-4

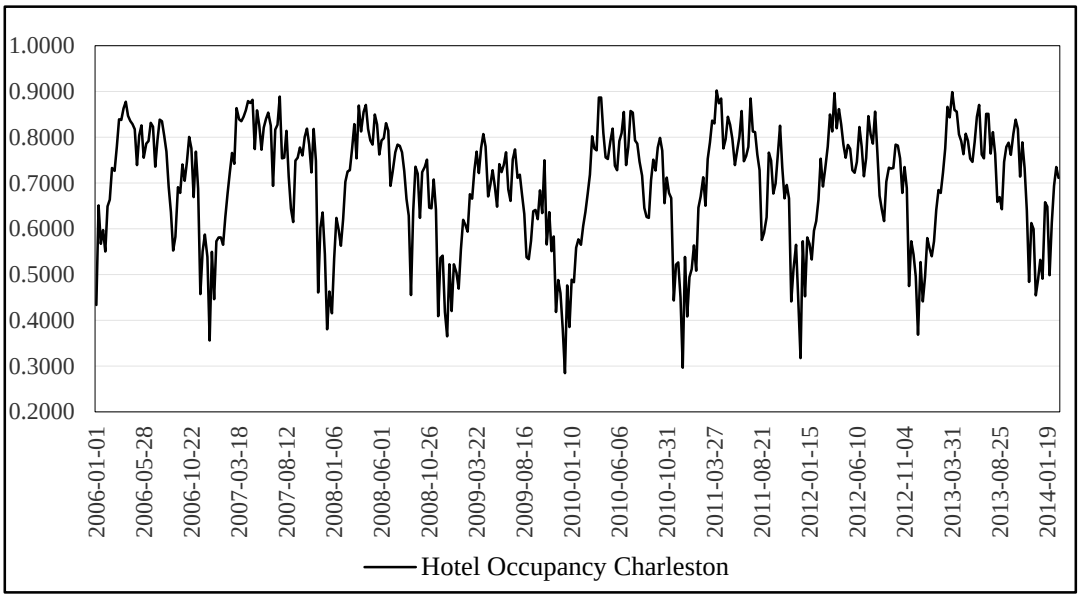


Figure 6. Weekly hotel occupancy in Charleston, SC

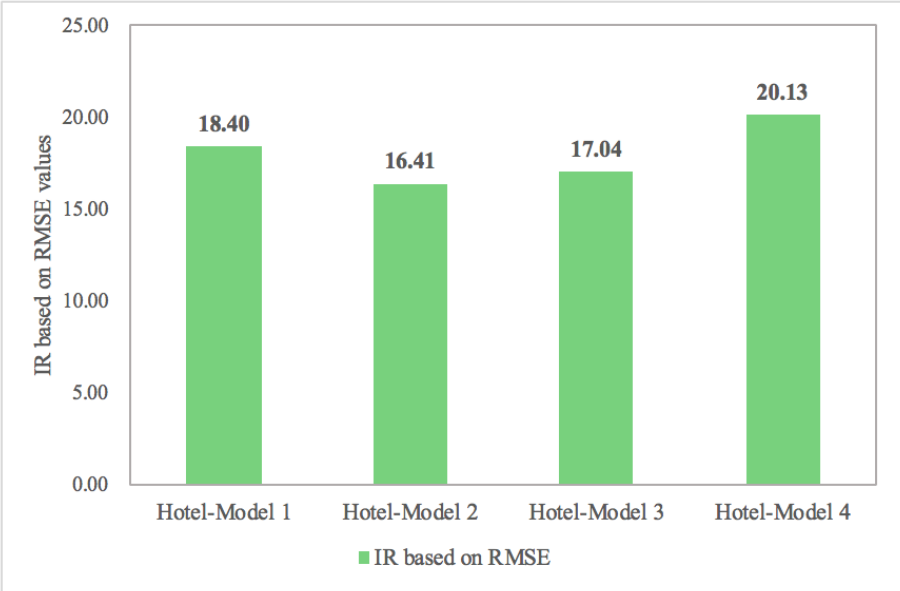


Figure 7. IR based on RMSE among four weekly hotel forecasting models