FORECASTING TOURISM DEMAND WITH MULTISOURCE BIG DATA

Hengyun Li, Ph.D.
School of Hotel and Tourism Management,
The Hong Kong Polytechnic University,
Hong Kong SAR, China
Email: neilhengyun.li@polyu.edu.hk

Mingming Hu*, Ph.D.
Business School, Guangxi University,
100# East of Daxue Road, Nanning 530004;
School of Hotel and Tourism Management,
The Hong Kong Polytechnic University, Hong Kong SAR, China.
Email: mingming.hu@gxu.edu.cn
Mobile Phone: +86 13877174602
*Corresponding Author.

Gang Li, Ph.D.
School of Hospitality & Tourism Management,
University of Surrey
Guildford, Surrey, GU2 7XH, UK
Email: g.li@surrey.ac.uk

Acknowledgments

This paper and research project (Project Account Code: 5-ZJLT) is funded by Research Grant of Hospitality and Tourism Research Centre (HTRC Grant) of the School of Hotel and Tourism Management, The Hong Kong Polytechnic University. This paper is also supported by the National Natural Science Foundation of China (71761001) and Hong Kong Scholars Program.

This is an Accepted Manuscript of an article published by Elsevier in International Journal of Hospitality Management in 2020. Available online:
https://doi.org/10.1016/j.annals.2020.102912

FORECASTING TOURISM DEMAND WITH MULTISOURCE BIG DATA

Abstract: Based on internet big data from multiple sources (i.e., the Baidu search engine and two online review platforms, Ctrip and Qunar), this study forecasts tourist arrivals to Mount Siguniang, China. Key findings of this empirical study indicate that (a) tourism demand forecasting based on internet big data from a search engine and online review platforms can significantly improve forecasting performance; (b) compared with tourism demand forecasting based on single-source data from a search engine, demand forecasting based on multisource big data from a search engine and online review platforms demonstrates better performance; and (c) compared with tourism demand forecasting based on online review data from a single platform, forecasting performance based on multiple platforms is significantly better.
Keywords: Tourism demand; Tourist attraction; Search engine; Online review; Multisource big data

Highlights

- This study forecasts weekly tourism arrivals to a national park in China.
- Internet big data from a search engine and online review platforms are employed.
- Findings suggest the superiority of multiple-source big data forecasting.
- Forecasting based on online review data from multiple platforms is preferred.

1. INTRODUCTION

Tourism demand forecasting plays an important role in the travel and tourism industry, and it provides important implications for destination policymakers and tourism practitioners (Colladon, Guardabascio, & Innarella, 2019). Predicting tourist arrivals is also important for the planning, operation, and management of tourist attractions (Huang, Zhang, & Ding, 2017). Specifically, Dergiades, Mavragani, and Pan (2018) stated that accurate tourism demand forecasting can benefit medium- to long-term marketing and tourism strategy development, pricing policies, investment plans and strategies, and allocation of limited resources. Given its importance, precise and timely tourism demand forecasting has become an increasingly popular topic in academic research.

Traditional tourism demand forecasting relies on structured statistical data published by governments. Yet forecasting is inherently limited by delayed and low-frequency publication of such data, leading to inaccurate predictions (Huang, Zhang, & Ding, 2017). Internet big data offer a valuable opportunity to provide timely tourism demand forecasting and to increase forecasting accuracy. These data can measure and monitor tourist behaviors and satisfaction in a timely manner while overcoming lags in traditional forecasting methods (Huang, Zhang, & Ding, 2017). Therefore, internet big data are effective supplements to traditional data sources (Choi & Varian, 2012; Wamba et al., 2015). Yang, Pan, and Song (2014) contended that internet big data can reveal tourists' preferences and their changes in real time in addition to providing high-frequency information (e.g., daily or weekly). Up-to-date information on tourist changes compensates for the limitations of tourism demand forecasting when using traditional data, as such methods often fail to forecast tourism demand accurately in cases of one-off events where data patterns change (Dergiades, Mavragani, & Pan, 2018).

Extensive research has applied internet big data, such as search engine data or website traffic data, to forecast tourism demand. Several empirical studies have demonstrated the usefulness of search query data in improving the forecasting of tourism demand (Bangwayo-Skeete & Skeete, 2015; Li, Chen, Wang, & Ming, 2018; Li, Pan, Law, & Huang, 2017; Sun, Wei, Tsui, & Wang, 2019), hotel room demand (Pan, Wu, & Song, 2012), and tourist attraction demand (Huang, Zhang, & Ding, 2017; Peng, Liu, Wang, & Gu, 2017). Apart from search query data, website traffic data have also been found to improve the forecasting accuracy of hotel demand in a destination (Pan & Yang, 2017; Yang, Pan, & Song, 2014).

Tourism businesses and destinations can also gain useful insight from content analysis of social media data, such as online reviews, and customers prefer to trust peer-supplied reviews rather than information from service providers (Xiang, Schwartz, Gerdes, & Uysal, 2015). Similarly, social media data can help practitioners anticipate rapid changes in tourists' preferences and popularity trends related to destinations and local attractions; such information can be gleaned from the number of online reviews and tourists' sentiments embedded within them. Some studies have revealed the usefulness of online reviews in forecasting product sales beyond tourism contexts (Dellarocas, Zhang, & Awad, 2007; Fan, Che, & Chen, 2017; Schneider & Gupta, 2016; Yu, Liu, Huang, & An, 2010). Accordingly, online reviews have been deemed highly important, with the potential to be incorporated into

tourism demand predictions (Colladon, Guardabascio, & Innarella, 2019).

Although previous studies have indicated that internet big data can greatly enhance tourism demand forecasting performance and offer valuable practical implications, several research gaps in tourism forecasting with such data should be addressed. First, most research has relied on volume-based search engine data or website traffic data for tourism demand forecasting; few studies have referred to volume- and sentiment-based social media data, which are much richer and can reflect tourists' attention and sentiments. Even so, volume-based data have their own shortcomings: a higher volume of website traffic does not necessarily reflect greater consumer interest in visiting a destination; in fact, the opposite may be true. For example, the Hong Kong protests in 2019 garnered increasing online attention, but the number of visitors to Hong Kong actually declined amidst safety concerns. Therefore, it would make sense to integrate volume-based and complementary sentiment-based variables when forecasting tourism demand. In particular, consumer-generated online reviews from online travel websites provide useful reflections of consumers' behaviors and satisfaction (Ye, Law, & Gu, 2009; Xiang, Schwartz, Gerdes, & Uysal, 2015), yet this type of data has yet to be employed to forecast tourism demand. Second, most prior studies considered internet big data from a single source, either from a search engine or the website of a specific destination marketing organization. However, few studies have investigated tourism demand forecasting performance by including big data from multiple sources in a single forecasting model. Overly narrow and insufficiently diverse data are major culprits of poor model forecasting; under such circumstances, models do not perform well in a variety of cases (Phillips et al., 2017). This limitation can be overcome by incorporating data from multiple, often complementary sources (Jia et al., 2016; Phillips et al., 2017; Pan & Yang, 2017). On this basis, this study will address the following research question: Can incorporating internet big data, including search query data and online review data, into a model improve forecasting accuracy over a model using only internet search query data?

2. LITERATURE REVIEW

2.1 Common Methods of Tourism Demand Forecasting

Common approaches to tourism demand forecasting consist of time series models, econometric models, and artificial intelligence (AI) models (Song & Li, 2008; Li, Pan, Law, & Huang, 2017). Classical time series models include the naïve model, exponential smoothing model, autoregressive–moving-average (ARMA) models, and structural time series model (Peng, Song, & Crouch, 2014). Although time series models offer distinct advantages in forecasting accuracy, they seldom consider the influencing factors of tourism demand, which may result in a loss of important information (Yang & Zhang, 2019). Econometric models are conducted on the basis of the relationship between a tourism demand variable and its influencing factors. In terms of econometric models, the most widely used models are autoregressive distributed lag model (ADLM), terror correction model, vector autoregressive model, and time-varying parameter (TVP) model, according to recent systematic reviews of tourism forecasting methods (Jiao & Chen, 2019; Wu, Song, & Shen, 2017). Moreover, some new developments of econometric techniques have emerged in recent tourism forecasting studies. For instance, a spatial autoregressive fixed-effect model has been introduced by Long et al. (2019), and Bayesian estimation has been incorporated into VAR models (Assaf et al., 2019; Gunter & Önder, 2016). AI models have also been used to forecast tourism demand, such as artificial neural networks, support vector machine (SVM), deep learning, and kernel extreme learning machines (Chen & Wang, 2007; Law, Li, Fong, & Han, 2019; Pai & Hong, 2005; Sun, Wei, Tsui, & Wang, 2019). Studies have shown that machine learning approaches can improve forecasting accuracy; however, they cannot offer theoretical explanations for associations between tourism demand data and other variables (Song & Li, 2008). Scholars have generally agreed that no single method can consistently outperform other methods on all occasions.

2.2 Current State of Tourism Forecasting Using Internet Big Data

Two primary types of internet big data have appeared in the tourism demand forecasting literature, namely search query data and website traffic data. The prevailing form of internet big data used in tourism forecasting is search query data, specifically generated from search engines such as Google and Baidu. Tourists use online search engines to plan trips; thus, search engine data can be used to forecast tourism demand (Dergiades, Mavragani, & Pan, 2018). Several studies have demonstrated that incorporating search engine data can improve tourism demand forecasting performance. Research has also found Baidu data to be more useful in depicting domestic tourism demand in China, whereas Google data outperforms Baidu in terms of forecasting international tourism demand (Yang, Pan, Evans, & Lv, 2015).

At the destination level, by using autoregressive mixed-data sampling (AR-MIDAS) models, Bangwayo-Skeete and Skeete (2015) noted that search query data related to ‘hotels and flights’ from Google could significantly improve the forecasting accuracy of tourist arrivals to Caribbean destinations. Moreover, AR-MIDAS was found to outperform the seasonal autoregressive integrated moving average and autoregressive (AR) approach in

tourism demand forecasting. Based on a dynamic linear model, Rivera (2016) discovered that search query data could improve the forecasting accuracy of the number of hotel nonresident registrations in Puerto Rico. Li and Law (2019) found that decomposed search engine data could be used to improve the accuracy of tourism demand forecasting from nine countries to Hong Kong. Furthermore, Dergiades, Mavragani, and Pan (2018) pointed out the usefulness of the corrected search query index (i.e., accounting for language and platform biases) in predicting international tourist arrivals to Cyprus; this approach also outperformed a forecasting model using an uncorrected search query index. However, tourism demand forecasting models may face challenges when several highly correlated search query indices are introduced into the model. To overcome this problem, Li, Pan, Law, and Huang (2017) proposed a procedure for calculating a composite search index by using a generalized dynamic factor model. They determined that the proposed method, along with the composite index, could significantly improve the accuracy of predicting tourist arrivals to Beijing. Similarly, Li, Chen, Wang, and Ming (2018) proposed a model named PCA-ADE-BPNN, a dimensional reduction algorithm, to predict tourism demand to Beijing and Hainan by using search query data from the Baidu index; their results indicated that the model outperformed other alternative models. AI models also appear helpful in processing high-dimensional data to overcome the high-correlation problem. Using internet search query data, Law, Li, Fong, and Han (2019) demonstrated that the deep learning approach could outperform other AI models, such as support vector regression and neural networks, in forecasting Macau tourist arrivals.

At the tourist attraction level, Huang, Zhang, and Ding (2017) used the Baidu index to forecast daily tourist arrivals to the Forbidden City in China, noting that incorporating search query data could significantly improve forecasting performance. Relatedly, Volchek, Liu, Song, and Buhalis (2019) demonstrated that search query data could enhance forecasts of visitor arrivals to five London museums. Different from the above studies, Peng, Liu, Wang, and Gu (2017) proposed an innovative method combining Hurst exponent (HE) and time difference correlation (TDC) analysis to select the most predictive search keywords. They found that the forecasting model based on keywords selected using the HE-TDC method showed better predictive ability when used to forecast visitor arrivals to the Jiuzhai Valley scenic area.

Other scholars have applied website traffic data to forecast tourism or hotel demand. Yang, Pan, and Song (2014) and Pan and Yang (2017) demonstrated that integrating web traffic data into the traditional time series model could significantly improve short-term forecasts of hotel room demand in Charleston, South Carolina. Moreover, Gunter and Önder (2016) used 10 website traffic indicators from Google Analytics to forecast tourism demand to Vienna, Austria; they found that incorporating these indicators could improve the forecasting accuracy for relatively longer horizons (h = 3, 6, and 12 months).

Regarding tourism demand forecasting using social media data, few studies have been conducted. Önder, Gunter, and Gindl (2019) applied the Likes data of posts on destination marketing organization (DMO) Facebook webpage as a predictor of tourism demand. Based on the restricted AR-MIDAS model and ADLM using Facebook Likes as an explanatory variable, it was found that the one-step-ahead mean forecasts of restricted AR-MIDAS model and ADLM outperform the benchmark naïve-1 model for cities of Graz and Vienna.

However, the opposite is true for Innsbruck and Salzburg. By using the sentiment of online news media coverage regarding a destination, Önder, Gunter, and Scharl (2019) forecasted tourist arrivals to Berlin, Brussels, Paris, and Vienna in Europe. The empirical results showed that the MIDAS model including news sentiment as an explanatory variable significantly outperforms the benchmark time-series models in terms of forecasting accuracy except for Vienna. Different from the above studies, Gunter, Önder, and Gindl (2019) investigated whether the combined Google Trends and Facebook Likes data can increase the tourism demand forecasting accuracy for four Austrian cities. However, the forecasting results are mixed among these four cities. For Salzburg, the ADLM including only Facebook Likes or both Likes and Google Trends data outperforms MIDAS and benchmark models in most cases. For Vienna, the MIDAS model including both Likes data and Google Trends generally shows the highest forecasting accuracy across forecasting horizons. However, for Graz and Innsbruck, the benchmark models outperform the ADLM and the MIDAS model.

2.3 Rationale for the Current Study

Although studies have demonstrated the benefits of different types of internet big data, few have considered multiple data sources in tourism demand forecasting (Pan & Yang, 2017). Pan and Yang (2017) investigated an optimal modeling technique to forecast weekly hotel occupancy based on combined big data sources, including search engine and website traffic data. Gunter, Önder, and Gindl (2019) investigated the potential of the combined Google Trends and Facebook Likes data in improving the accuracy of tourism demand forecasting. Their study provided preliminary evidence of the advantages of integrating multiple sources of big data into a forecasting model. Their findings thus substantiated the potential merits of integrating multiple data sources rather than relying on a single big data source.

In addition, data from a single source or platform may limit the stability and generalization of a model's forecasting performance; that is, forecasting may be successful in one context but fail in others (Phillips, Dowling, Shaffer, Hodas, & Volkova, 2017). One important reason for this model generalization problem is the mismatch between the training data in a model and the data that will be used for forecasting; in other words, training data tend to be too narrow in scope, which limits the forecasting model's ability to perform well in a variety of situations (Phillips, Dowling, Shaffer, Hodas, & Volkova, 2017). One way to overcome this issue is to draw data from multiple internet big data sources or platforms. This strategy has been applied successfully in other areas, such as demographic forecasting, where the relationships between social media behavior and user demographics differ by platform. Jia et al. (2016) and Song et al. (2016) argued that combining information from multiple internet platforms can increase the robustness of a forecasting model, which is especially beneficial when data are complementary.

Additionally, Pan and Yang (2017) revealed the limitations of multisource big data in tourism demand forecasting if the data are highly correlated. Their study indicated that including search query and web traffic data only reduced the mean absolute percentage error (MAPE) marginally compared with a forecasting model with only one data source. Given this limitation in big data forecasting, it would be particularly beneficial to use multisource big data that are more diverse and complementary to increase tourism demand forecasting

accuracy. Therefore, other big data sources that are distinct from search query or web traffic data, such as internet big data from social media, could potentially yield more valuable predictions.

In our context, search query data are unique from social media data. Searches are usually conducted in private and by a much larger population compared to social media discussions. However, search query data suffer from several pitfalls. For example, search query data are not as rich as social media data, as they can only indicate tourists' level of interest in a tourism product or destination but have limited ability to reflect tourist sentiment. Conversely, social media data are more detailed and can reflect tourists' intentions and sentiments, although these data are subject to insufficient and skewed representativeness of the user population and potential intentional manipulation. A common practice in predictive research to mitigate flaws in imperfect data involves supplementing the data with additional, non-overlapping, and meaningful information that does not suffer from the issues the investigator aims to alleviate (Geva, Oestreicher-Singer, Efron, & Shimshoni, 2017). Social media data, as an additional data source, may therefore complement search query data well in a forecasting exercise. The preceding evidence underpins our rationale for this study, which aims to forecast tourism demand by using online big data from multiple sources, including a search engine and social media online review platforms.

3. METHODOLOGY

We propose an integrated framework (see Figure 1) to incorporate search query and online review data into tourism demand forecasting. This framework includes four steps: 1) data collection, 2) data processing and variable calculation, 3) model specification, and 4) model estimation and forecasting performance evaluation. In the first step, we collected three types of data: weekly tourist arrival data from a tourist attraction's official website, search query data from Baidu's search engine, and online review data from Ctrip and Qunar. In the second step, we calculated three types of variables based on internet big data collected during Step 1, including the weekly search volumes of different keywords in Baidu's search engine, the weekly review volume, and the weekly average review rating. In the third step, we established a few models to test the role of multisource big data in improving tourism demand forecasting: (a) time series benchmark models, including seasonal Naïve (SNAIVE), Exponential Smoothing State Space (ETS) and autoregressive integrated moving average (ARIMA) models; (b) forecasting models with search query data, including an ARIMA model taking search query data as explanatory variables (ARIMAX), SVM, and random forest (RF); and (c) forecasting models (ARIMAX, SVM, and RF) including search query and online review data. Regarding our reasons for selecting the above forecasting models, earlier studies revealed the superiority of these models in accurate forecasting using high-frequency data (Cui, Gallino, Moreno, & Zhang, 2018; Geva, Oestreicher-Singer, Efron, & Shimshoni, 2017; Pan, Wu, & Song, 2012). In the fourth step, we estimated the model and evaluated its forecasting performance based on the mean absolute error (MAE), root mean square error (RMSE), MAPE, and Diebold-Mariano (DM) test (Harvey, Leybourne, & Newbold, 1997).

[*Insert Figure 1 Here*]

3.1 Data Description

We focused on this study on demand data for Mount Siguniang, a national park in China. Mount Siguniang is in Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province. The park was named a UNESCO Heritage Site as part of Sichuan Giant Panda Sanctuaries in 2006. There are four peaks in the park. The highest peak is the Yaomei Peak, at 6,250 m above sea level, which is a famous climbing destination; the other three peaks are popular hiking destinations for tourists.

We measured tourism demand by tourist arrivals. Weekly tourist arrivals to Mount Siguniang from January 2, 2017 to July 14, 2019 (132 weeks in total) were collected from Mount Siguniang's official website (https://www.sgns.cn/news/number) (see Figure 2). To forecast tourist arrivals to Mount Siguniang, we focused on two types of internet big data, namely search query and online review data.

[*Insert Figure 2 Here*]

First, search query volume data were gathered from Baidu's search engine. Search engines enable tourists to explore destinations and organize their trips. Their search behaviors are recorded throughout this process. We chose Baidu as our search query data source given its prominence in China, where this search engine occupies more than 70% of market share (Cui, 2019). Following the keyword selection method of Li et al. (2017) and keyword inclusion in the Baidu index, we chose eight Chinese-language keywords related to Mount Siguniang: 'Mount Siguniang Travel Guide (四姑娘山攻略)', 'Mount Siguniang's weather (四姑娘山天气)', 'Mount Siguniang's altitude (四姑娘山海拔)', 'Where is Mount Siguniang (四姑娘山在哪里)', 'Tourist attractions in Mount Siguniang (四姑娘山景区)', 'Tickets to Mount Siguniang (四姑娘山门票)', 'Travel at Mount Siguniang (四姑娘山旅游)', and 'Hotel at Mount Siguniang (四姑娘山住宿)'. We used the Baidu search index to measure each keyword's search volume. The daily Baidu search index for each keywork was collected from its website (http://index.baidu.com/) by self-compiled python crawler tools and then aggregated to weekly search volumes.

Second, online review data were collected from online review platforms. Online review data include two dimensions, specifically weekly review volume and weekly average review rating. Ctrip (www.ctrip.com) and Qunar (www.qunar.com) are well-known online travel service platforms in China. After visitation, verified Ctrip or Qunar consumers can post tourist attraction reviews online. There were 1,367 reviews on Qunar and 678 on Ctrip from January 2, 2017 to July 14, 2019. Self-compiled python crawler tools were utilized to collect these reviews. To combine online reviews from both sites, all reviews along with their ratings were mixed, and we retained only one if two reviews were identical in their review rating and textual content. Ultimately, we collected 1,645 unique online reviews and ratings from both sites, which were then ranked by their time stamp and aggregated into two variables on a weekly basis: 1) weekly review volume; and 2) weekly average review rating. The average review rating was taken as a reflection of tourists' satisfaction level, whereas the number of reviews conveyed the popularity of a given tourist attraction. Data descriptions appear in Table 1.

[*Insert Table 1 Here*]

3.2 Modeling Strategy

The full dataset spanned 132 weeks. We conducted separate forecasting from 1 week to 12 weeks ahead. Data on the first 107 weeks of tourist arrivals, search queries, and online reviews were used for our initial model estimation, after which the 1- to 12-weeks-ahead forecasts were generated. The estimation sub-sample was then extended by one week each time, and rolling forecasts of up to 12 weeks ahead were generated until all available data had been incorporated. The accuracy across 12 forecasting horizons was evaluated thereafter.

To explore the role of internet big data from multiple sources in tourism demand forecasting, seven models (see Table 2) were constructed. SNAIVE, ETS and ARIMA model are time series models, which depends on historical tourism demand data only. ARIMAX$_1$ is

an ARIMA model that takes search query data as explanatory variables. SNAIVE, ETS, ARIMA represent basic benchmark models. $ARIMAX_2$, $ARIMAX_3$ and $ARIMAX_4$ are ARIMAX models including search query and online review data as explanatory variables: $ARIMAX_2$ contains combined online review data from Qunar and Ctrip, $ARIMAX_3$ includes online review data from Qunar, and $ARIMAX_4$ consists of review data from Ctrip.

[*Insert Table 2 Here*]

As noted earlier, scholars have yet to reach a consensus on an optimal method for modeling tourism demand when using internet big data. Therefore, two AI models, SVM and RF, were constructed and estimated for comparison with ARIMAX in terms of forecasting performance. Two scale-dependent errors and a percentage error were used to evaluate forecasting accuracy. The equations for calculating MAE, RMSE, and MAPE are as follows:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (1)$$

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

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (3)$$

where y_i denotes the actual number of tourist arrivals; $\hat{y}_i$ denotes the predicted number of tourist arrivals; and n denotes the number of forecasts for evaluation.

3.3 Forecasting Models

3.3.1 ARIMA/ARIMAX. ARIMA is a classical time series stochastic process model, which contains autoregression (AR), moving averages (MA), and a difference component (Cho, 2003). ARIMA has been widely applied in many research areas, such as electricity price forecasting (Contreras, Espinola, Nogales, & Conejo, 2003), stock price forecasting (Pai & Lin, 2005), and tourism demand forecasting (Claveria & Torra, 2014). This model's general form (Pankratz, 2009) is as follows:

$$\left(1 - \sum_{i=1}^{p}\phi_i L^i\right)(1-L)^{\mathrm{d}}Y_t = \left(1 + \sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \quad (4)$$

where $\left(1 - \sum_{i=1}^{p}\phi_i L^i\right)$ is the AR component; $\left(1 + \sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$ is the MA component; and $(1-L)^{\mathrm{d}}$ is the d times difference.

When explanatory variables (i.e., search query and/or online review variables in this study) are included in the model, the ARIMA model is known as ARIMAX. It takes the form of

$$\left(1 - \sum_{i=1}^{p}\phi_i L^i\right)(1-L)^{\mathrm{d}}Y_t = \sum_{k=1}^{u}\sum_{i=1}^{r}\eta_{ki}L^i X_{kt} + \left(1 + \sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \quad (5)$$

where X_{kt} is/are the explanatory variable(s); L^i denotes ith lags; and η_{ki} is the coefficient for $L^i X_{kt}$.

The parameters p and q in Equation 5 are determined by an auto-correlation function test and partial auto-correlation function test (Box, Jenkins, & Reinsel, 1994). The parameter d is determined by using the unit root test (Dickey & Fuller, 1979). Explanatory variables and their lags can be determined using the Akaike information criterion index (Akaike, 1974).

3.3.2 Support Vector Machine (SVM). As a classical machine learning algorithm, SVM was developed to forecast tourism demand in this study. SVM can map data x into a high-dimensional feature space through a nonlinear mapping function (Akın, 2015). SVM then generates a unique and globally optimal solution by solving a linearly constrained quadratic programming problem, which is resistant to over-fitting. According to Huang, Nakamori, and Wang (2005) and Pai and Lin (2005), the SVM model is introduced as follows:

Suppose a set of training vectors ($G = \{(x_i, y_i), i = 1,2, \dots, N; x_i \in R^n; y_i = -1 \; or \; 1\}$) belongs to two separate classes. A hyperplane $\mathrm{w}^{\mathrm{T}}\varphi(x) + b = 0$ is used to classify the $x_i (i = 1,2, \dots, N)$, which should satisfy the following conditions:

$$\mathrm{w}^{\mathrm{T}}\varphi(x_i) + b \geq 1 \quad \text{if } y_i = 1,$$

$$\mathrm{w}^{\mathrm{T}}\varphi(x_i) + b \leq -1 \text{ if } y_i = -1. \tag{6}$$

Equivalently,

$$y_i[\mathrm{w}^{\mathrm{T}}\varphi(x_i) + b] \geq 1, \mathrm{i} = 1,2, \dots, \mathrm{N} \tag{7}$$

where $\varphi: R^n \to R^m$ is the feature mapping the input space to a usually high-dimensional feature space.

Thus, the distance between two margins is $2/\|w\|$, and the optimal parameter w^*, b^* of the hyperplane can be determined by solving $min\frac{1}{2}\|w\|^2$. In the training process, the classification function determined by w^*, b^* is given such that

$$f(x) = \mathrm{Sign}(w^{*T}x + b^*) \tag{8}$$

The regression problem involves seeking the fitting function: $f(x) = w \cdot \varphi(x) + b$. When applying SVM in regression, the goal is to seek a hyperplane to fit the given points. The minimization principle of structural risk is also taken as the goal to construct a mathematical model:

$$\min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\vartheta_{\mathrm{i}} + \theta_i)$$

$$\text{S.T.} \quad [\mathrm{w}^{\mathrm{T}}\varphi(x_i) + b] - y_i \leq \varepsilon + \vartheta_i, i = 1,2, \dots, N$$

$$y_i - [\mathrm{w}^{\mathrm{T}}\varphi(x_i) + b] \leq \varepsilon + \theta_i, i = 1,2, \dots, N$$

$$\vartheta_i, \theta_i \geq 0, i = 1,2, \dots, N \tag{9}$$

where the constant C is the cost controlling the training error; ϑ_i, θ_i denote imported errors from the training set; and ε is an insensitive loss function, representing the permitted training

loss.

A Lagrange multiplier method can be used to solve the above model and generate the non-linear mapping function:

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \beta_i)K(x_i, x) + b \quad (10)$$

where $K(x_i, x)$ is a Kernel function, which should satisfy Mercer's condition (Vapnik, 1995); N is the number of training samples; α_i and β_i are the Lagrange coefficients; and b is a constant. α_i, β_i, and b are each generated from the Lagrange multiplier method.

In forecasting models based on SVM, lagged weekly tourist arrivals and all big data indices (i.e., search query and online review variables) were incorporated as input variables, and current weekly tourist arrivals were taken as output variables. These input and output variables together were used to construct training samples (x). After training, the estimated Equation (10) were used to generate forecasts. Package "kernlab" in R was used for training and forecasting (Karatzoglou, Smola, Hornik & Karatzoglou, 2019). For a detailed introduction to SVM, please refer to Campbell (2001).

3.3.3 Random Forest (RF). RF, proposed by Breiman (2001), is an ensemble method that randomly produces a diverse pool of individual regression systems (Bernard et al., 2009). This approach combines the ideas of classification and regression tree (CART) (Breiman, Friedman, Olshen, & Stone, 1984) and bagging (Breiman, 1996). RF extends CART by introducing the bagging method, which improves the stability and accuracy of learning algorithms (Khaidem, Saha, & Dey, 2016). According to Khaidem et al. (2016) and Tyralis and Papacharalampous (2017), the main steps of RF are as follows:

Step 1: Select the fitting set. A group of observations S_s is selected randomly from the training dataset u for fitting. All big data indices and other influencing factors are incorporated as input variables to construct training dataset u.

Step 2: Plant trees. Randomly generate M trees. M groups of subspaces in the feature space $(\theta_1, \theta_2, \ldots, \theta_M)$ are randomly created. For a function f, the predicted value at u is denoted by

$$f_s(u; \theta_j, S_s). \quad (11)$$

The random feature variable θ_j is used to resample the fitting set to grow individual trees and select successive directions for splitting.

Step 3: Split node. A parent node splits into two daughter nodes. The splitting decision is intended to reduce impurity or gain as much information as possible. The information gain due to a split can be calculated by

$$\Delta g(N) = g(N) - P_L g(N_L) - P_R g(N_R) \quad (12)$$

where P_L is the proportion of the population of the left daughter node; P_R is the proportion of the population of the right daughter node; and $g(N)$ is the Gini impurity measure in node N.

The splitting process stops when each cell contains fewer than node size. After the tree stops growing, predicted values $f_s(u; \theta_j, S_s)$ can be generated by the tree.

Step 4: Fusion prediction value created by all trees. When RF is used for forecasting, the average value of prediction values generated by all trees is the final output. For a detailed

explanation of RF, please refer to Breiman (2001).

The package ‘randomForest’ in R can be used to train RF and generate forecasts (RColorBrewer & Liaw, 2018). In our study, lagged weekly tourist arrivals and all big data indices (i.e., search query and online review variables) were incorporated as input variables, and current weekly tourist arrivals were taken as output variables. These input and output variables together were used to construct training samples (x). After training the RF model, tourist arrivals were forecasted by putting new values of input variables into the trained model.

4. RESULTS

Two groups of comparisons were conducted to answer our research question (see Figure 3). The first comparison group was used to test whether incorporating internet big data (i.e., search query data and online review data) into a single forecasting model could improve forecasting accuracy; the second comparison group was used to test whether combining online review data from Ctrip and Qunar into one forecasting model could improve the forecasting accuracy.

[*Insert Figure 3 Here*]

To test whether incorporating search query data and online review data into one forecasting model would improve forecasting accuracy compared to a benchmark time series model and a model using only internet search query data, SNAIVE, ETS, ARIMA, $ARIMAX_1$ and $ARIMAX_2$ models were constructed and estimated. We took SNAIVE, ETS, ARIMA and $ARIMAX_1$ models as benchmarks. SNAIVE, ETS and ARIMA models are time series models, while $ARIMAX_1$ is an ARIMAX model with search query volume as a leading indicator. $ARIMAX_2$ is an ARIMAX model with search query data from Baidu and online review data from Qunar and Ctrip as leading indicators. The DM test was conducted to evaluate the statistical significance of the forecasting accuracy improvement of one model against another. Taking the improvement of $ARIMAX_1$ (compared to ARIMA) on MAE as an example, the equation for calculating improvement is as follows:

$$\text{Improvement} = \frac{\text{MAE(ARIMA)} - \text{MAE}(\text{ARIMAX}_1)}{\text{MAE(ARIMA)}} \times 100\% \quad (13)$$

Forecasting results, improvements and DM test results compared with benchmarks are presented in Table 3. First, compared with the benchmarks (the SNAIVE, ETS and ARIMA models), $ARIMAX_1$ (with search query volume is a leading indicator) improved the forecasting accuracy when forecasting 1 to 6 weeks ahead; however, $ARIMAX_1$ performed worse when forecasting 9 and 12 weeks ahead. This indicates that the search query volume is likely to be effective only for short-term tourist arrival forecasting in the case of Mount Siguniang. However, compared with SNAIVE, ETS and ARIMA, $ARIMAX_2$ (including search query volume and online review variables as leading indicators) enhanced the forecasting accuracy consistently from 1 to 12 weeks ahead. The result of the DM test shows that the improvements of $ARIMAX_2$ are consistently significant, statistically speaking, when comparing with SNAIVE and ETS, while the improvements are significant only in forecasting 1 to 3 weeks ahead when comparing with ARIMA. Second, when taking $ARIMAX_1$ as a benchmark, $ARIMAX_2$ also improved the forecasting accuracy significantly and consistently across all 12 forecasting horizons: MAE improved from 27.87% to 54.96%; RMSE improved from 28.28% to 49.80%; and MAPE improved from 22.03% to 62.35%.

The DM test result further shows that these improvements are statistically significant at least at the 10% significance level across all horizons. These results together demonstrate the superiority of multisource big data forecasting for Mount Siguniang compared with forecasting using single-source big data or traditional time series models.

[*Insert Table 3 Here*]

Online travel service platforms provide an array of travel-related information, and tourists can post online reviews for consumed travel products and services. In China, Ctrip and Qunar are well-known online travel service markets with relatively distinct business focuses (Liu, 2015). Ctrip functions more as an online tourism agent that interacts directly with suppliers and users. Qunar currently generates most of its income through advertisements, although it has begun to shift toward serving as an online travel agent (Liu, 2015). These different focuses correspond to different prices and services, thus attracting unique groups of consumers. Therefore, the information included on these two online travel service platforms could vary. To test the roles of online reviews from these platforms in tourism demand forecasting, we compared $ARIMAX_2$, $ARIMAX_3$ and $ARIMAX_4$. Online review data in $ARIMAX_2$ were taken from Qunar and Ctrip, whereas online review data for $ARIMAX_3$ and $ARIMAX_4$ were from either Qunar or Ctrip, respectively. These models' forecasting performance is summarized in Table 4.

According to the MAE, RMSE, MAPE values in Table 4, $ARIMAX_2$ consistently outperformed $ARIMAX_3$ and $ARIMAX_4$, and these improvements are statistically significant according to the DM test. Therefore, combined online review data from the two chosen websites performed better in tourism demand forecasting compared to review data from either site alone. Upon comparing the forecasting performance of $ARIMAX_3$ and $ARIMAX_4$, $ARIMAX_4$ was found to perform better; that is, online review data from Ctrip resulted in better forecasts than data from Qunar. This reveals that for tourist arrival forecasting of Mount Siguniang, the prediction performance based on both Ctrip and Qunar platforms is significantly better than the prediction using a single online review data platform. This implies that the two platforms captured different segments of tourists to Mount Siguniang, who shared different experiences on social media. The reviews on Ctrip and Qunar complemented each other and both sources of reviews influenced the visit intention of future tourists.

[*Insert Table 4 Here*]

Next, we considered whether the above results (from ARIMA/ARIMAX models) could be obtained using other AI models. We performed a robustness check by applying two AI models, SVM and RF, and estimating two groups of models. We took search query

volume as an explanatory variable in the first group of models; we considered search query and online reviews on both platforms as explanatory variables in the second group. ARIMAX, SVM, and RF were used to forecast weekly tourist arrivals to Mount Siguniang. The forecasting accuracy of these models is shown in Table 5 along with the improvements of the second group over the first. The positive role of online reviews in enhanced tourism demand forecasting was evident for SVM and RF in short-term forecasting (i.e., 1, 2, 3 and 6 weeks ahead), seen from the improvements of MAE, RMSE and MAPE values, although the improvements were not as significant as the case of $ARIMAX_2$ against $ARIMA_1$. When forecasting arrivals of 9 and 12 weeks ahead, the positive role of online review data in forecasting performance improvement could not be consistently verified by any AI models in terms of MAE, RMSE, MAPE or DM test results. Essentially, the positive role of online review data in enhancing tourism demand forecasting appeared more effective for short-term forecasting.

[*Insert Table 5 Here*]

5. CONCLUSIONS AND IMPLICATIONS

Internet big data have revolutionized how tourism demand is forecasted (Yang, Pan, & Song, 2014; Volchek, Liu, Song, & Buhalis, 2018). Based on the case of a national park in China, the empirical results of this study revealed that compared with the benchmark model without any internet big data variables, tourism demand forecasting incorporating internet big data from a search engine and online review platforms could significantly improve forecasting performance. Moreover, we found that compared with tourism demand forecasting based on single-source big data from a search engine, tourism demand forecasting based on multisource big data from a search engine and online review platforms elicited better short-term forecasting performance. Moreover, combining the same type of big data over different platforms could increase the model's forecasting performance. Specifically, compared with tourism demand forecasting based on online review data from a single platform (either Ctrip or Qunar), forecasting performance was significantly enhanced when using data from multiple platforms (Ctrip and Qunar).

The contributions of this study are multifold. First, we made an initial attempt to incorporate both tourist attention and tourist sentiment variables into a tourism demand forecasting model; The vast majority of previous studies only considered either tourist attention based on search query or website traffic volume frequency analysis or tourist sentiment in a demand forecasting system. Second, our study is one of the earliest research to enhance demand forecasting performance for visitor attractions/tourist destinations based on both search engine and social media data (i.e., online review data). The study by Gunter, Önder, and Gindl (2019) is the only best known exception, which forecasts the destination tourist arrivals using both Google search query data and Likes data of DMO Facebook webpage. Indices from multiple sources of internet big data offer a more comprehensive overview and stronger theoretical support for improving forecasting performance. Third, this study is among the first to apply social media online review data from multiple platforms (i.e., Ctrip and Qunar) in tourism demand forecasting. Each of these online review platforms is geared toward a certain population; a major challenge of using a single review platform involves potential sample bias. Tourism demand forecasting that incorporates user-generated reviews from multiple online platforms should therefore provide better population coverage and result in higher forecasting accuracy.

This study also unveils practical implications for managers of tourism attractions and destinations. First, indices based on internet big data from multiple sources can offer richer information about tourists' interests and preferences, thus enabling more accurate demand forecasting for tourism attractions and destinations. This novel approach highlights the importance of multisource online big data and could revolutionize tourism forecasting systems in the long run. Second, based on nearly real-time and high-frequency forecasting, tourist attraction operators can adjust daily demand predictions as needed and achieve revenue management objectives by applying dynamic pricing strategies and appropriate staff scheduling. Moreover, authorities can use short-term visitor forecasts to support crowd management and to increase a destination's competitiveness in the long run.

This study has a few limitations that lend themselves to further investigation. First,

the constructed tourism demand forecasting models tended to incorporate only internet big data and a lagged tourism demand variable without considering other important influencing factors in tourism demand. Therefore, subsequent studies could examine whether integrating internet big data from multiple sources alongside traditional influencing factors of tourism demand into one forecasting model might improve forecasting accuracy. Second, we only considered one type of social media data, namely online review data, to forecast tourism demand. Therefore, future studies can extend social media data by incorporating additional types of user-generated social media data, such as information from Facebook, microblogs, and internet discussion forums, to improve the accuracy of tourism demand forecasting. Third, as with most tourism forecasting studies, our research is case study based. The choice of a single case was restricted by the data availability for comparable attraction sites. Therefore, the findings of this study should be not generalized. The purpose of our study is to illustrate the potential benefit of using multisource big data in tourism forecasting. By nature, it is explorative. To gain more generalizable conclusion, further investigations are necessary by incorporating multiple cases where data are available.

6. REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. In Selected Papers of Hirotugu Akaike (pp. 215-222). New York, NY: Springer.

Akın, & Melda. (2015). A novel approach to model selection in tourism demand modelling. *Tourism Management, 48,* 64-72.

Assaf, A. G., Li, G., Song, H., & Tsionas, M. G. (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., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46*, 454-464.

Bernard, S., Heutte, L., & Adam, S. (2009). Influence of hyperparameters on random forest accuracy. In International Workshop on Multiple Classifier Systems (pp. 171-180). Berlin, Heidelberg: Springer.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). Time series analysis, forecasting and control. Englewood Cliffs, NJ: Prentice-Hall.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 26*, 123-140

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth.

Campbell, C. (2001). An introduction to kernel methods. *Studies in Fuzziness and Soft Computing, 66*, 155-192.

Chen, K. Y., & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management, 28*(1), 215-226.

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management, 24*(3), 323-330.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2-9.

Claveria, O., & Torra, S. (2014). Forecasting tourism demand to Catalonia: Neural networks vs. time series models. *Economic Modelling, 36*, 220-228.

Colladon, A. F., Guardabascio, B., & Innarella, R. (2019). Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. *Decision Support Systems*, *123*, 113075.

Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. *IEEE Transactions on Power Systems, 18*(3), 1014-1020.

Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management*, *27*(10), 1749-1769.

Cui, Y. (2019, January 28). 2019 Ranking of search engines in China: Baidu, Shenma, Sougou, 360's market share. Retrieved from https://www.seoxiehui.cn/article-107466-1.html.

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23-45.

Dergiades, T., Mavragani, E., & Pan, B. (2018). Google Trends and tourists' arrivals: Emerging biases and proposed corrections. *Tourism Management*, *66*, 108-120.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366a), 427-431.

Fan, Z. P., Che, Y. J., & Chen, Z. Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. *Journal of Business Research*, *74*, 90-100.

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using forum and search data for sales prediction of high-involvement products. *MIS Quarterly*, *41*(1),65-82.

Goh, C., & Law, R. (2003). Incorporating the rough sets theory into travel demand analysis. *Tourism Management*, 24(5), 511–517.

Gunter, U., & Önder, I. (2016). Forecasting city arrivals with Google Analytics. *Annals of Tourism Research*, *61*, 199-212.

Gunter, U., & Önder, I. (2016). Forecasting city arrivals with Google Analytics. *Annals of Tourism Research*, *61*, 199-212.

Gunter, U., Önder, I., & Gindl, S. (2019). Exploring the predictive ability of LIKES of posts on the Facebook pages of four major city DMOs in Austria. *Tourism Economics*, *25*(3), 375-401.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, *13*(2), 281-291.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research, 32*(10), 2513-2522.

Huang, X., Zhang, L., & Ding, Y. (2017). The Baidu Index: Uses in predicting tourism flows–A case study of the Forbidden City. *Tourism Management*, *58*, 301-306.

Jia, Y., Song, X., Zhou, J., Liu, L., Nie, L., & Rosenblum, D. S. (2016, February). Fusing social networks with deep learning for volunteerism tendency prediction. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Jiao, E. X., & Chen, J. L. (2019). Tourism forecasting: a review of methodological developments over the last decade. *Tourism Economics*, *25*(3), 469-492.

Karatzoglou, A., Smola, A., Hornik, K., & Karatzoglou, M. A. (2019). Package 'kernlab'. Technical report, CRAN, 03 2016.

Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.

Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research*, *75*, 410-423.

Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*, *68*, 116-126.

Li, X., & Law, R. (2019). Forecasting Tourism Demand with Decomposed Search Cycles. *Journal of Travel Research*. https://doi.org/10.1177/0047287518824158

Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57-66.

Liu, R. (2015, October 27). How will Ctrip and Qunar conquer 70% of China's tourism market and 70 million users? Retrieved from https://medium.com/@actallchinatech/how-will-

ctrip-and-qunar-conquer-70-of-china-s-tourism-market-and-70-million-users-f94d914d00f8.

Long, W., Liu, C., & Song, H. (2019). Pooling in tourism demand forecasting. *Journal of Travel Research*, *58*(7), 1161–74.

Önder, I., Gunter, U., & Gindl, S. (2019). Utilizing Facebook Statistics in Tourism Demand Modeling and Destination Marketing. *Journal of Travel Research*, 0047287519835969.

Önder, I., Gunter, U., & Scharl, A. (2019). Forecasting tourist arrivals with the help of web sentiment: A mixed-frequency modeling approach for big data. *Tourism Analysis*, *24*(4), 437-452.

Pai, P. F., & Hong, W. C. (2005). An improved neural network model in forecasting arrivals. *Annals of Tourism Research, 32,* 1138-1141.

Pai, P. F., & Lin, C. S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega, 33*(6), 497-505.

Pan, B., & Yang, Y. (2017). Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research*, *56*(7), 957-970.

Pan, B., Wu, D. C., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, *3*(3), 196-210.

Pankratz, A. (2009). Forecasting with univariate Box-Jenkins models: Concepts and cases. New York, NY: John Wiley & Sons.

Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management*, *45*, 181-193.

Peng, G., Liu, Y., Wang, J., & Gu, J. (2017). Analysis of the prediction capability of web search data based on the HE-TDC method–prediction of the volume of daily tourism visitors. *Journal of Systems Science and Systems Engineering*, *26*(2), 163-182.

Phillips, L., Dowling, C., Shaffer, K., Hodas, N., & Volkova, S. (2017). Using social media to predict the future: a systematic literature review. *arXiv preprint arXiv:1706.06134*.

RColorBrewer, S., & Liaw, M. A. (2018). Package 'randomForest'. *University of California, Berkeley: Berkeley, CA, USA.*

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management*, *57*, 12-20.

Schneider, M. J., & Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, *32*(2), 243-256.

Song, H., & Li, G., (2008). Tourism demand modelling and forecasting-A review of recent research, *Tourism Management, 29*(2), 203-220.

Song, X., Ming, Z. Y., Nie, L., Zhao, Y. L., & Chua, T. S. (2016). Volunteerism tendency prediction via harvesting multiple social networks. *ACM Transactions on Information Systems (TOIS)*, *34*(2), 10.

Sun, S., Wei, Y., Tsui, K.-L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, 70, 1-10.

Tyralis, H., & Papacharalampous, G. (2017). Variable selection in time series forecasting using random forests. *Algorithms, 10*(4), 114.

Vapnik, V. (1995). The nature of statistical learning theory. New York, NY: Springer.

Volchek, K., Liu, A., Song, H., & Buhalis, D. (2019). Forecasting tourist arrivals at attractions: Search engine empowered methodologies. *Tourism Economics*, *25*(3), 425-447.

Wamba, F. S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234-246.

Wu, D. C., Song, H., & Shen, S. (2017). New development in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, *29*(1), 507 – 529.

Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International Journal of Hospitality Management*, *44*, 120-130.

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, *46*, 386-397.

Yang, Y., & Zhang, H. (2019). Spatial-temporal forecasting of tourism demand. *Annals of Tourism Research*, 75, 106-119.

Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research*, *53*(4), 433-447.

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*(1), 180-182.

Yu, X., Liu, Y., Huang, X., & An, A. (2010). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, *24*(4), 720-734.

Table 1.

Data description

Data Category	Variable	Data source	Min.	Max.	Mean	Median	Std. Dev.
Tourism Demand	Weekly tourist arrivals	Government website	1052	73322	11069	8490	11296
Search Query	Mount Siguniang travel guide	Baidu index	1263	9115	3296	2728	1712
	Mount Siguniang's weather	Baidu index	1551	13848	4740	4694	2396
	Mount Siguniang's altitude	Baidu index	650	5542	1997	1755	943
	Where is Mount Siguniang	Baidu index	773	6128	2230	1872	1030
	Tourist attractions in Mount Siguniang	Baidu index	0	1326	548	506	330
	Tickets to Mount Siguniang	Baidu index	373	2786	1062	988	401
	Travel at Mount Siguniang	Baidu index	0	1264	568	558	294
	Hotel at Mount Siguniang	Baidu index	0	1632	629	629	345
Online Reviews	Weekly review volume	Qunar.com &Ctrip.com	0	92	11.70	7	14.51
	Weekly average review rating	Qunar.com &Ctrip.com	0	5	4.56	4.87	1.04

Table 2.

Variables included in the forecasting models

Model	Historical Series	Search Query	Online Review (Qunar & Ctrip)	Online Review (Qunar)	Online Review (Ctrip)
SNAIVE	√				
ETS	√				
ARIMA	√				
$ARIMAX_1$	√	√			
$ARIMAX_2$	√	√	√		
$ARIMAX_3$	√	√		√	
$ARIMAX_4$	√	√			√

Table 3.

Forecasting accuracy and improvements

Horizon	Model	MAE	RMSE	MAPE	$ARIMAX_1$ vs. others				$ARIMAX_2$ vs. others			
					MAE	RMSE	MAPE	DM test	MAE	RMSE	MAPE	DM test
1	SNAIVE	5469.44	9169.83	0.7681	**19.23%**	**45.02%**	-2.14%	-1.3924*	**51.82%**	**62.02%**	**42.26%**	-1.7596**
1	ETS	4978.05	8729.55	0.5975	**11.25%**	**42.24%**	-31.31%	-1.3247*	**47.07%**	**60.10%**	**25.77%**	-1.6216*
1	ARIMA	4712.17	7471.53	0.8963	**6.25%**	**32.52%**	**12.47%**	-1.0283	**44.08%**	**53.38%**	**50.52%**	-1.4678*
1	$ARIMAX_1$	4417.9	5042.02	0.7845					**40.36%**	**30.92%**	**43.47%**	-2.4889**
1	$ARIMAX_2$	***2634.94***	***3482.92***	***0.4435***								
2	SNAIVE	5647.75	9355.76	0.7711	**21.64%**	**46.50%**	**10.87%**	-1.4468*	**51.80%**	**61.63%**	**45.06%**	-1.7480**
2	ETS	5500.36	9725.84	0.9024	**19.54%**	**48.54%**	**23.84%**	-1.4000*	**50.51%**	**63.09%**	**53.05%**	-1.6362*
2	ARIMA	4919.62	7663.66	1.1213	**10.05%**	**34.69%**	**38.71%**	-1.1084	**44.66%**	**53.16%**	**62.21%**	-1.4703*
2	$ARIMAX_1$	4425.36	5005.07	0.6873					**38.48%**	**28.28%**	**38.35%**	-2.2750**
2	$ARIMAX_2$	***2722.27***	***3589.68***	***0.4237***								
3	SNAIVE	5813.91	9549.4	0.7681	**14.42%**	**39.94%**	-0.84%	-1.2601	**50.90%**	**61.54%**	**42.73%**	-1.7442**
3	ETS	5449.29	8525.45	0.8464	**8.69%**	**32.73%**	**8.49%**	-1.0049	**47.61%**	**56.92%**	**48.02%**	-1.5190*
3	ARIMA	5042.05	7540.81	1.1393	**1.32%**	**23.95%**	**32.01%**	-0.7659	**43.38%**	**51.30%**	**61.38%**	-1.4095*
3	$ARIMAX_1$	4975.54	5735.02	0.7746					**42.63%**	**35.97%**	**43.20%**	-2.5613***
3	$ARIMAX_2$	***2854.7***	***3672.34***	***0.44***								
6	SNAIVE	6355.45	10366.6	0.6648	**29.27%**	**39.37%**	**11.15%**	-1.0154	**48.98%**	**62.05%**	**30.65%**	-1.4711*
6	ETS	5891.81	10220.69	0.5757	**23.70%**	**38.51%**	-2.60%	-1.0313	**44.96%**	**61.51%**	**19.92%**	-1.4868*
6	ARIMA	4993.93	7500.44	1.0889	**9.98%**	**16.20%**	**45.76%**	-0.4377	**35.07%**	**47.55%**	**57.66%**	-1.2815
6	$ARIMAX_1$	4495.37	6285.17	0.5907					**27.87%**	**37.40%**	**21.95%**	-1.4012*
6	$ARIMAX_2$	***3242.69***	***3934.29***	***0.461***								
9	SNAIVE	6272.82	7805.35	0.5706	**3.67%**	**5.68%**	-25.75%	-0.1959	**48.21%**	**47.28%**	**39.03%**	-1.7159*
9	ETS	7131.21	10992.79	0.6406	**15.26%**	**33.03%**	-12.01%	-0.8726	**54.44%**	**62.57%**	**45.69%**	-1.4410*
9	ARIMA	4446.91	7117.07	0.4734	-35.89%	-3.44%	-51.57%	0.084	**26.94%**	**42.19%**	**26.51%**	-0.9319
9	$ARIMAX_1$	6042.77	7362.08	0.7175					**46.24%**	**44.11%**	**51.51%**	-2.0291**
9	$ARIMAX_2$	***3248.73***	***4114.59***	***0.3479***								
12	SNAIVE	6974.86	8408.29	0.5979	-6.83%	-3.33%	-42.19%	0.1099	**51.88%**	**48.13%**	**46.59%**	-1.6884*
12	ETS	6700.67	8258.37	0.563	-11.20%	-5.21%	-50.99%	0.1649	**49.91%**	**47.18%**	**43.29%**	-1.5784*
12	ARIMA	4186.55	7163.5	0.3465	-77.98%	-21.29%	-145.32%	0.4821	**19.83%**	**39.11%**	**7.85%**	-0.7882
12	$ARIMAX_1$	7451.25	8688.34	0.8501					**54.96%**	**49.80%**	**62.44%**	-2.4456**
12	$ARIMAX_2$	***3356.19***	***4361.77***	***0.3193***								

Note: The Bold and Italic numbers are the best performance compared with other models; The Bold numbers are the percentages of improvements. ***, ** and * denote statistical significance at the 1%, 5% and 10% levels, respectively. A negative DM test value implies that the former model provides better predictions than the latter model in each comparison.

Table 4.
Forecasting accuracy and improvements compared with $ARIMAX_3$ and $ARIMAX_4$

Horizon	Model	MAE	RMSE	MAPE	$ARIMAX_2$ vs. others			
					MAE	RMSE	MAPE	DM test
1	$ARIMAX_2$	***2634.94***	***3482.92***	***0.4435***				
1	$ARIMAX_3$	3983.63	4856.23	0.7936	**33.86%**	**28.28%**	**44.11%**	-1.6598**
1	$ARIMAX_4$	3220.20	3794.03	0.6186	**18.17%**	**8.20%**	**28.31%**	-0.6491*
2	$ARIMAX_2$	***2722.27***	***3589.68***	***0.4237***				
2	$ARIMAX_3$	3915.12	4813.33	0.5988	**30.47%**	**25.42%**	**29.25%**	-1.4484**
2	$ARIMAX_4$	2942.46	3764.63	0.4707	**7.48%**	**4.65%**	**9.98%**	-0.2569*
3	$ARIMAX_2$	***2854.70***	***3672.34***	***0.4400***				
3	$ARIMAX_3$	4975.54	5735.02	0.7746	**42.63%**	**35.97%**	**43.20%**	-1.9190***
3	$ARIMAX_4$	3460.78	4264.68	0.5506	**17.51%**	**13.89%**	**20.10%**	-0.7934**
6	$ARIMAX_2$	***3242.69***	***3934.29***	***0.4610***				
6	$ARIMAX_3$	4601.06	6361.64	0.5872	**29.52%**	**38.16%**	**21.49%**	-1.4279*
6	$ARIMAX_4$	3321.52	4401.93	0.4768	**2.37%**	**10.62%**	**3.32%**	-0.4584*
9	$ARIMAX_2$	***3248.73***	***4114.59***	***0.3479***				
9	$ARIMAX_3$	6011.57	7356.94	0.7120	**45.96%**	**44.07%**	**51.13%**	-2.0418**
9	$ARIMAX_4$	4152.98	5261.76	0.5112	**21.77%**	**21.80%**	**31.93%**	-0.9846**
12	$ARIMAX_2$	***3356.19***	***4361.77***	***0.3193***				
12	$ARIMAX_3$	7396.61	8464.10	0.8275	**54.63%**	**48.47%**	**61.41%**	-2.4577**
12	$ARIMAX_4$	4916.53	6093.61	0.5615	**31.74%**	**28.42%**	**43.13%**	-1.0120**

Note: The Bold and Italic numbers are the best performance compared with other models; The Bold numbers are the percentages of improvements. ***, ** and * denote statistical significance at the 1%, 5% and 10% levels, respectively. A negative DM test value implies that $ARIMAX_2$provides better predictions than $ARIMAX_3$ or $ARIMAX_4$.

Table 5.

Forecasting accuracy with or without online review data

Horizon	Model	X=Search Query			X=Search Query & Online Review			Improvement: X=Search Query & Online Review vs. X=Search Query			
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	DM test
1	ARIMAX	4417.90	5042.02	0.7845	2634.94	3482.92	0.4435	**40.36%**	**30.92%**	**43.47%**	-2.4889**
1	SVM	3668.12	6827.18	0.4566	3319.52	6705.03	0.3346	**9.50%**	**1.79%**	**26.72%**	-1.2599
1	RF	3491.26	4563.27	0.5305	3281.94	4427.07	0.5088	**6.00%**	**2.98%**	**4.09%**	-1.0611
2	ARIMAX	4425.36	5005.07	0.6873	2722.27	3589.68	0.4237	**38.48%**	**28.28%**	**38.35%**	-2.2750**
2	SVM	3982.58	5772.73	0.5245	3576.58	5616.62	0.4125	**10.19%**	**2.70%**	**21.35%**	-1.0216
2	RF	3964.48	4850.50	0.5992	3749.16	4638.22	0.5759	**5.43%**	**4.38%**	**3.89%**	-1.8106**
3	ARIMAX	4975.54	5735.02	0.7746	2854.70	3672.34	0.4400	**42.63%**	**35.97%**	**43.20%**	-2.5613***
3	SVM	4166.57	6010.44	0.4619	3779.26	5772.91	0.4263	**9.30%**	**3.95%**	**7.71%**	-1.1529
3	RF	4407.79	4962.61	0.6629	4330.76	4837.81	0.6600	**1.75%**	**2.51%**	**0.44%**	-0.8578
6	ARIMAX	4495.37	6285.17	0.5907	3242.69	3934.29	0.4610	**27.87%**	**37.40%**	**21.95%**	-1.4012*
6	SVM	3911.90	5797.33	0.4056	3723.20	5734.97	0.4055	**4.82%**	**1.08%**	**0.03%**	-0.6540
6	RF	4681.88	5379.46	0.5972	4585.53	5266.71	0.5970	**2.06%**	**2.10%**	**0.04%**	-0.5597
9	ARIMAX	6042.77	7362.08	0.7175	3248.73	4114.59	0.3479	**46.24%**	**44.11%**	**51.51%**	-2.0291**
9	SVM	4510.53	6281.62	0.3917	4360.47	6357.81	0.3949	**3.33%**	-1.21%	-0.82%	0.2368
9	RF	5741.35	6423.36	0.6221	5687.13	6287.95	0.6262	**0.94%**	**2.11%**	-0.67%	-0.4169
12	ARIMAX	7451.25	8688.34	0.8501	3356.19	4361.77	0.3193	**54.96%**	**49.80%**	**62.44%**	-2.4456**
12	SVM	4168.64	6256.62	0.3404	4016.29	6374.04	0.3434	**3.65%**	-1.88%	-0.90%	0.2969
12	RF	6025.03	6534.52	0.6346	6149.76	6652.61	0.6459	-2.07%	-1.81%	-1.78%	0.4696

Note: The Bold numbers are the percentages of improvements. ***, ** and * denote statistical significance at the 1%, 5% and 10% levels, respectively. A negative DM test value implies that the model with both research query and online review variables outperforms the competing model with the search query variable only.

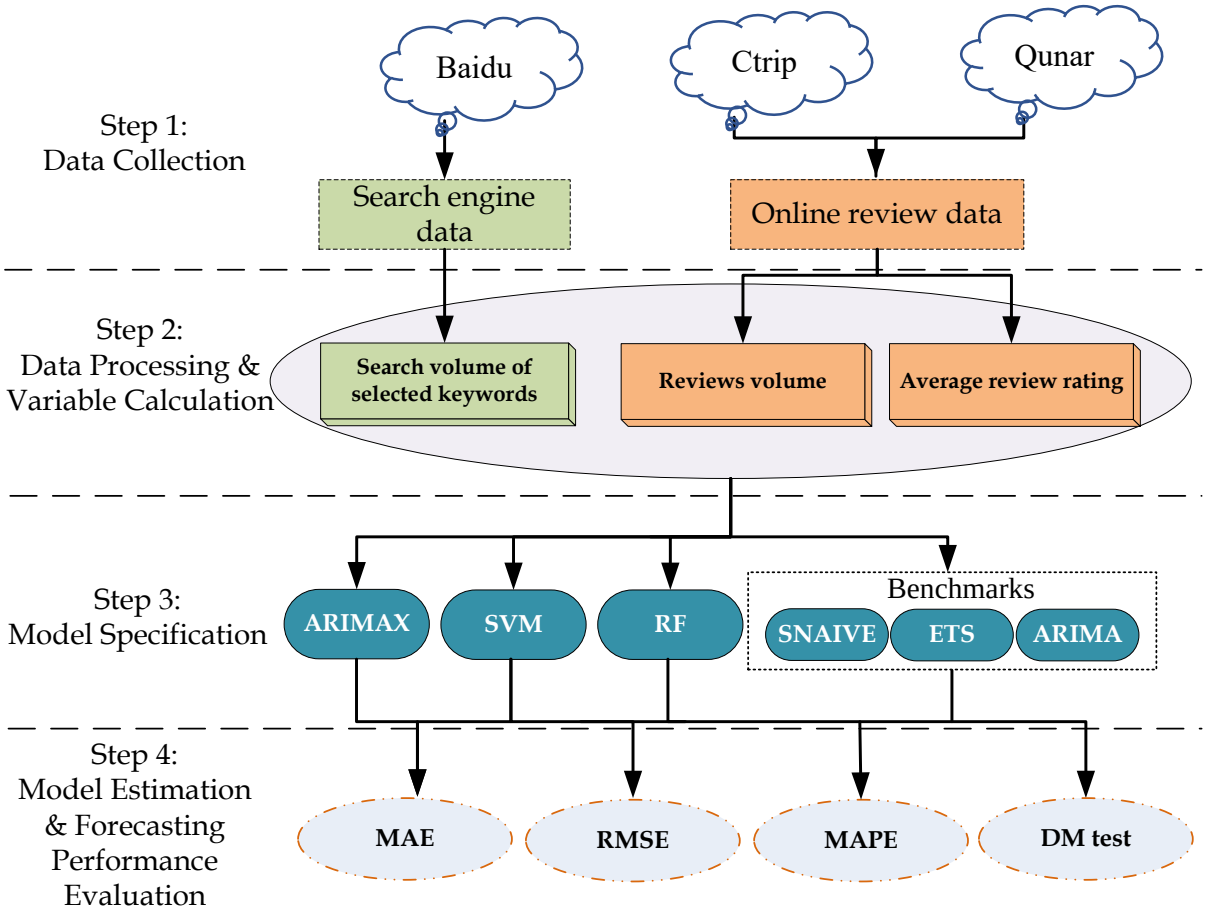

Figure 1. Forecasting framework

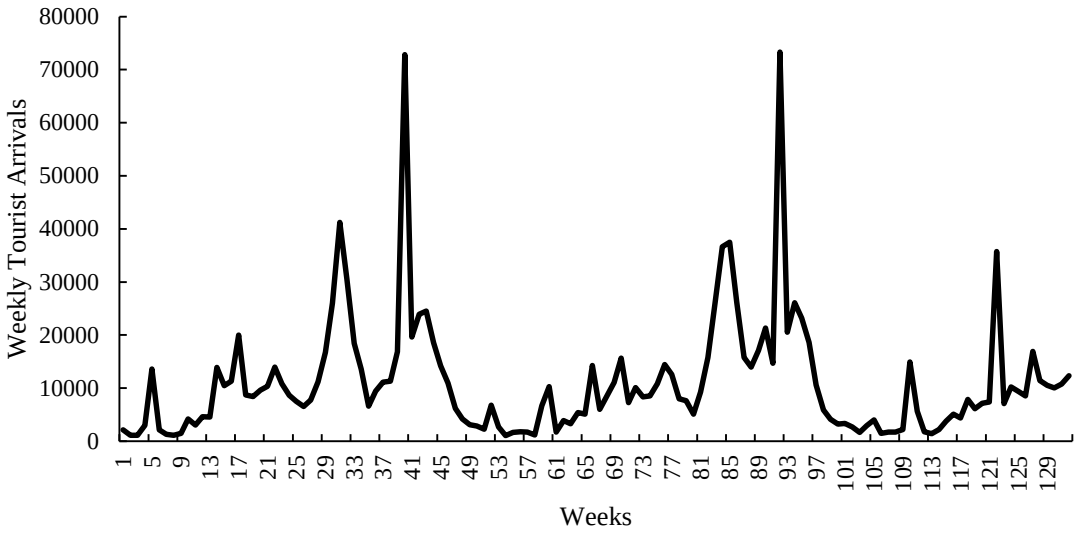

Figure 2. Weekly tourist arrivals at Mount Siguniang

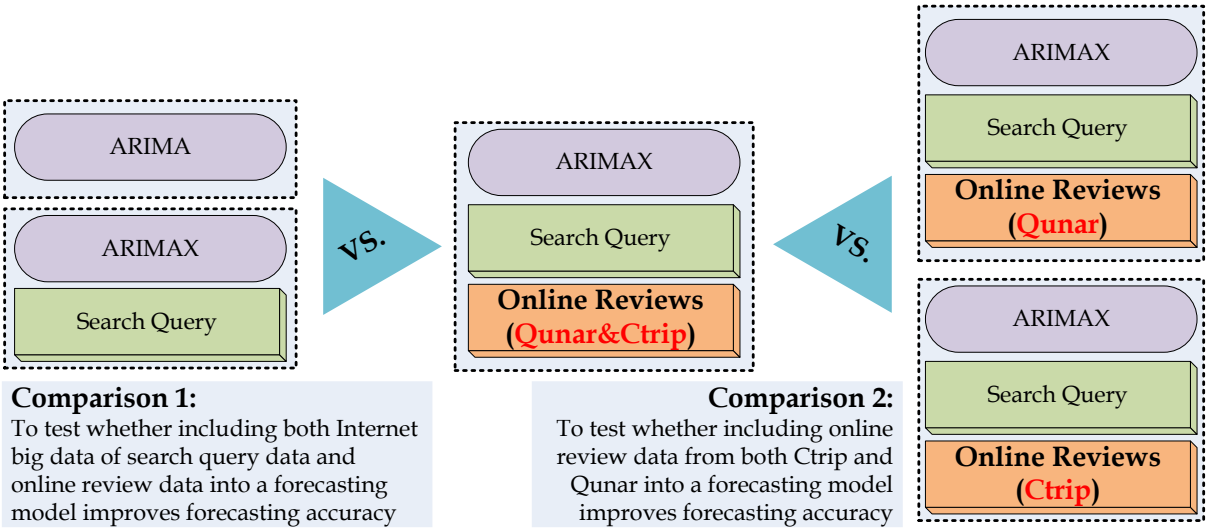

Figure 3. Two groups of comparison