Forecasting tourism demand with composite search index:

Article · April 2017
DOI: 10.1016/j.tourman.2016.07.005

4 authors, including:

Bing Pan
Pennsylvania State University
97 PUBLICATIONS 3,735 CITATIONS

Some of the authors of this publication are also working on these related projects:

The impact of contextual cues on response rate, conversion rate, and destination preference in travel surveys View project

All content following this page was uploaded by Bing Pan on 25 August 2016.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.
Forecasting Tourism Demand with Composite Search Index

Xin Li, Ph.D.,
Institute of Tourism, Beijing Union University
Beijing 100101, China
Email: leexin111@163.com

Bing Pan, Ph.D.
Associate Professor
Department of Hospitality and Tourism Management
School of Business, College of Charleston, Charleston, USA
Email: bingpan@gmail.com
Visiting Professor
School of Tourism and Environmental Sciences
Shaanxi Normal University, Xi’an, China

Rob Law, Ph.D.
Professor
School of Hotel and Tourism Management
The Hong Kong Polytechnic University, Kowloon, Hong Kong
Email: rob.law@polyu.edu.hk

Xiankai Huang, Ph.D.
Professor
Beijing Open University
Beijing 100081, China
Email: xiankai@buu.edu.cn

Submitted Exclusively to Tourism Management
January, 2016

1 This research was partially supported by grants from the National Natural Science Foundation of China (NSFC No. 71373023 and NSFC No. 41428101) and the Talents Program in Beijing Union University (RK100201509).
Forecasting Tourism Demand with Composite Search Index

Abstract
Researchers have adopted online data such as search query volumes to forecast tourism demand for a destination, including tourist volumes and hotel occupancy. However, the massive yet highly correlated query data pose challenges when researchers attempt to include them in the forecasting model. We propose a framework and procedure for creating a composite search index adopted in a generalized dynamic factor model (GDFM) to forecast tourist demand in a destination. This research empirically tests the framework in predicting monthly Beijing tourist volumes. Findings suggest that the proposed method improves the forecast accuracy of monthly Beijing tourist volumes compared with two benchmark models: a time series model and a model with an index created by principal component analysis. The method demonstrates the combination of composite search index and a GDFM in accurately forecasting tourism demand.

Keywords: tourism demand forecast, big data analytics, search query data, generalized dynamic factor model, composite search index
1 Introduction

The advancements of the information technology have brought massive amount of big data generated by users, including search queries, social media mentions, mobile device positions, and others. (Mayer-Schonberger & Cukier 2013). In particular, search query data provide valuable information about tourists’ intention, interests, and opinions. Tourists use search engines to obtain weather and traffic information and to plan their routes by searching for hotels, attractions, travel guides, and other tourists’ opinions (Fesenmaier, Xiang, Pan, & Law, 2011). Search query data, including content and volume data, are especially valuable to researchers. They can capture tourists’ attention to travel destinations and can be extremely useful in accurately forecasting tourist volumes in a destination. The abundant search trends data are favorable sources for tourism forecasting in the Big Data era. However, they also bring challenges in the modeling process of tourism forecasts.

In particular, in forecasting tourist volumes of one destination with search trends data, one needs to collect tourism-related keywords, obtain their search trends data, select appropriate data to construct an aggregated index, and construct the econometric models. The major challenges are keyword selection and search data aggregation. Keyword selection has received significant attention from researchers. For example, Brynjolfsson, Geva, and Reichman (2015) proposed a crowd-squared method to select keywords. They prompted individuals through an online interface to produce word associations and verified that this method performed efficiently in the keyword selection task. In addition, the selected search trends data series are usually multi-dimensional, and thus considering the variables that should be added in the forecast models is important in improving prediction accuracy.

Compared with keyword selection methods, the process of index aggregation has received limited attention (Brynjolfsson, Geva, & Reichman, 2015). This step is generally conducted through three main approaches: (1) incorporating the keywords directly into the models, (2) extracting the index using the principal component analysis (PCA), and (3) index aggregation from multiple variables (Yang, Pan, Evans, & Lv, 2015). Although these approaches could predict variables more accurately than their benchmark models, they are still not optimal. First, multicollinearity or overfitting problems may occur when the dimensions of variables are high (Varian, 2014). In particular, out-of-sample forecasts may fail even when in-sample forecasts perform well. Second, a large amount of the original information will be lost if the data series is weighed equally in aggregating an index from multiple keywords. The forecast accuracy may be reduced because of incomplete information.

This study aims to propose a feasible variable selection method in forecasting tourist volumes with search trends data. The approach follows two rules. First, it should acquire one representative and meaningful index reflecting the dynamic correlation among all search trends data series. Second, the new method should be able to deal with a large number of search data series. As a result, a
generalized dynamic factor model (GDFM) is adopted to incorporate many keyword variables. An advantage of GDFM is its ability to process high-dimensional data and to use the composite index (Amstad & Potter, 2009). GDFM is commonly adopted in the analysis of economic or financial cycles (Amstad & Potter, 2009), but it is seldom used in the fields of tourism forecasting. We applied our proposed methodology to predict tourist volumes in Beijing, which is one of the most renowned travel destinations in China. By collecting specific search trends data from Baidu including tourism-related keywords (“dining,” “lodging,” “trip,” “traveling,” “shopping,” and “recreation”), this study empirically tested the method in the forecasting of monthly Beijing tourist volumes from January 2011 to August 2015. The empirical results demonstrate that our method is superior over the benchmark models of an autoregressive model and a model with PCA as predictor.

This paper proceeds as follows. Section II briefly reviews the relevant literature. Section III proposes the framework with integrated index construction. Section IV presents our empirical study and research findings. Finally, Section V concludes by discussing the study’s contributions and implications for future research.

2 Literature Review

In this section, we first review the current studies on tourism demand forecasting. Second, we focus on big data forecasting with search trends data, including the major techniques in the keyword and variable selection. We also introduce the generalized dynamic factor models along with their applications. Third, we address the research gap at the end of this section.

2.1 Tourism demand forecasting: data and techniques

Tourism demand forecasting is a well-established research area, and it has attracted many studies in the tourism and hospitality field. Song and Li (2008) gave a detailed literature review on tourist demand forecasting methods and techniques in recent decades. The commonly adopted forecasting techniques are time series, econometric models, artificial intelligence approaches, and hybrid methods.

Time series models predict tourist arrivals by considering past patterns. Many studies used time series models to analyze and forecast tourism demand (Gunter & Önder, 2015; Guizzardi & Stacchini, 2015; Akın, 2015; Chu 2009; Chu 2008; Athanasopoulos & Hyndman, 2008). The most popular ones are autoregressive moving average models (Song & Li, 2008). Econometric models explore the causal relationship between tourist arrivals and influencing factors, which are especially useful when a correlational relationship exists (Wong et al., 2007; Wong, Song, & Chon, 2006; Song & Witt, 2006; Song & Witt, 2000; Song, Witt, & Jensen, 2003). Artificial intelligence methods adopt neural networks and support vector machines to model the nonlinear data series (Palmer, Montano, & Sese, 2006; Hadavandi et al., 2015; Pai & Hong, 2005; Pai et al., 2006; Palmer et al., 2006). Some studies have proposed hybrid forecasting by combining econometric and data mining techniques (Sun et al., 2016; Pai et al., 2014). In addition, methods such as meta-analysis and
singular spectrum analysis are also used in the modeling and forecasting of tourist arrivals (Hassani et al., 2015; Peng, Song, & Crouch, 2014).

2.2 Big data analytics in tourism research

Big data analytics has become increasingly important in both the academic and the business communities over the past two decades (Chen, Chiang, & Storey, 2012). Travelers’ decision making is intrinsically complicated and multi-dimensional, and it includes many aspects such as selecting destinations, reserving hotels, planning trips, and other activities. The new data sources generated by users based on Internet technology (search engines or social media platforms) have become popular in studying travelers’ decision making and behavior.

Some extant literature has attempted to introduce user-generated content and big data analytics in tourism-related research. With big data sources, tourist arrivals or hotel sales can be forecasted more accurately. Choi and Varian (2012) investigated the predictive ability of search trends data series in travel destinations planning. By using keyword search volume data from Google, they increased the prediction accuracy for Hong Kong tourist arrivals from several countries such as the United States, Canada, Great Britain, Germany, and other countries. Yang, Pan, and Song (2013) predicted hotel demand by combining traditional econometric models with web traffic volumes and demonstrated the use of web volumes in predicting hotel occupancy in a tourist destination.

In addition, these search engine and social media data sources can also help improve customer service, user experience, and satisfaction (Pan, Litvin, & Goldman, 2006). Ye, Law, and Gu (2009) examined the effects of online consumer-generated reviews on hotel room sales. The data were collected from the largest travel website in China. Their research findings indicated a significant relationship between online reviews and the business performance of hotels. Li, Law, Vu, Rong, and Zhao (2015) used online reviews data from TripAdvisor to identify emergent hotel features of interests to international travelers. Their research findings helped hotel managers to gain insights into travelers’ interests and to better understand rapid changes in tourist preferences. Ghose, Ipeirotis, and Li (2012) proposed a hotel demand estimation model by combining US hotel reservation data from Travelocity with various social media sources. They applied big data analytics such as text mining, image classification, and social geo-tagging to generate a new ranking system and to provide customers with best-value hotels. Wohlfarth et al. (2011) collected Internet-based data and used the descriptive characteristics of flights and text mining to predict travel price changes. The findings help customers to decide when to purchase the best-value tickets. Xiang, Schwartz, Gerdes Jr., and Uysal (2015) exploited the manner in which big data analytics understands the relationship between hotel guest experiences and satisfaction. Big data and text analytics can discover customers’ behavior and represent their experiences.

Compared with traditional data sources in tourism research, big data analytics provides a large amount of data without sampling bias. With these new data sources, the academia and industries can better understand consumer behavior in the travel and hospitality fields.
2.3 Forecasting with search trends data

Researchers have adopted search volume data to forecast many social and economic activities. The forecasted variables include unemployment (Askitas & Zimmermann, 2009), consumption levels (Swallow & Labbé, 2013; Vosen & Schmidt, 2011), consumer prices (Choi & Varian, 2012), housing prices (Wu & Brynjolfsson, 2015), and stock prices (Da, Engelberg, & Gao, 2011).

A more recent trend in tourism demand forecasting is forecasting with search trends data from Google and Baidu. For example, Pan, Wu, and Song (2012) used five travel-related Google search volume data to predict hotel room demand in an autoregressive moving average with an exogenous model. Bangwayo-Skeete and Skeete (2015) used Google search and a mixed data sampling model to improve the forecasting performance of tourist arrivals. Yang et al. (2015) used Baidu and Google search trends data to predict Chinese tourist flows by using autoregressive moving average models and evaluated the performances of the data of the two search queries.

In general, researchers use three main data modeling techniques for search volume data. When the number of relevant search queries is small, researchers include volume data directly in the model (Vosen & Schmidt, 2011; McLaren, 2011). To accomplish the task, researchers use the autoregressive moving average with exogenous variables to construct the forecasting model. When the number of search query data is large, keeping all the variables in the model poses problems because of potential multicollinearity and overfitting issues in the model estimation (Varian, 2014). Thus, constructing indices from a large number of search query data is a feasible solution. Researchers can use PCA to construct a search index (Li, Shang, Wang, & Ma, 2015). In addition, researchers can aggregate data using data shift and summation in consideration of the lag orders of different types search query data (Yang, Pan, Evans, & Lv, 2015). Thus, the index is the linear combination of original search query data series.

The latter two methods create a search index to be included in the forecasting model. In general, the PCA index is the first factor extracted from the search query data, and the last method is a linear combination of the partials of the original data. However, both indices may fail to comprehensively represent the dynamics among all search queries. First, the PCA index is created with the reduction of the originally high dimensions, thus resulting in information loss in the process. Second, the last method using shift and summation linearly combines the search queries, and the aggregated index cannot address the dynamic correlations among search queries. For example, search query data are dynamically correlated with one another and can be determined by some potentially common factors such as holidays. Therefore, researchers need a new index construction method to process the multi-dimensional search queries effectively. The new index should be comprehensive and should reflect the most relevant information in search queries. It should also depict the common components of search queries to cover the lead and lag information of the search data series.

2.4 Generalized dynamic factor model
One model seems to be an ideal candidate for modeling multi-dimensional search data. Forni et al. (2000) proposed GDFM by extending the dynamic factor models. Let \( \{ x_{ij}, i = 1, \ldots, n, t = 1, \ldots, T \} \) be the set of observed variables. Each variable can be modeled as the sum of its common component \( \chi_i \) and an idiosyncratic component \( \xi_i \). The common components are driven by a \( q \)-dimensional vector of common factors \( f_1(t), f_2(t), \ldots, f_q(t) \). The model is noted as follows:

\[
x_i = \chi_i + \xi_i = B(L)f_i + \xi_i \quad \text{Eq. 3.1}
\]

\[
\chi_i = b_{1i}(L)f_{1t} + b_{2i}(L)f_{2t} + \cdots + b_{qi}(L)f_{qt} \quad \text{Eq. 3.2}
\]

where \( L \) is the lag operator, \( B(L) = b_j(L), i = 1, 2, \ldots, n, j = 1, 2, \ldots, q \) is the set of time-varying factor loadings, and \( q \) indicates the number of commonly dynamic factors. Forni et al. (2000) suggested that \( q \) is determined through the variance contribution of each component. If the variance contribution rate of the first \( i-1 \) components diverges and component \( i \) begins to converge, then \( q \) is set to \( i-1 \). The figure of variance contribution rate is used to determine the value of \( q \). The estimation details of GDFM are found in Forni et al. (2000) and Forni et al. (2005).

The model has two distinct superiorities in analyzing data with a large number of variables. First, the model can dynamically update parameters, and thus it can deal with typically dynamic questions. Second, GDFM allows for cross-correlation among idiosyncratic components. Specifically, GDFM can generate a coincident index to represent the common states of the observed variables. In the existing literature, not only can GDFM reflect the business cycle or inflation, but it can also forecast the dependent variable with cycles. Recently, GDFM has been widely used in the modeling and forecasting of business cycles (Christophe, 2006), underlying inflation gauges (Amstad & Potter, 2009), and other economic indicators (Forni et al., 2003).

However, the existing literature seldom uses this method to model search trends data in the field of tourism. The field of tourism is dynamically correlated, and the relevant industrial sectors such as restaurants, hotels, and travel agencies are also closely correlated with one another. Therefore, GDFM is an ideal candidate for dealing with multiple tourism-related search trends data series.

### 3 Methodology

#### 3.1 Forecasting framework with search trends data

After introducing GDFM, we propose a forecasting framework with search trends data. This framework describes a modeling process that starts from selecting user-generated data sources, processing data series, and constructing index to building forecasting models and evaluating their performances. We present the following six steps of this modeling framework:

(I) Select user-generated data sources. Depending on country and culture, different search sources can be used. For example, several search engines in China such as Baidu and Google can provide
search services for millions of users. In this empirical study, Baidu search trends data are selected as the major source because Baidu holds the highest market share compared with other search engines. According to the 35th report issued by the China Internet Network Information Center (CNNIC, 2015), the Google search engine is less popular in China, with a penetration rate of 27.4%, which is significantly less than 92.1% of Baidu in 2014.

(II) Select keywords. The purpose of this step is to define tourism-related keywords. This step intends to follow the mental process of a Chinese traveler who plans to visit a certain city. The traveler first needs to choose the destination; thereafter, he/she may want to search details about the food, specialties, traffic, weather, hotels, and insights into the place. Thereafter, we define all of these aspects, namely, dining, shopping, recreation, lodging, traffic, and insights, as the major factors that travelers consider in their upcoming trips. The search keywords are selected according to these aspects.

(III) Process data. This step intends to clean the search trends data series. The data are standardized to 0–100 to eliminate the influences of scales. The value of 100 indicates the highest search query volume, and 0 represents the lowest. Moreover, if the data present an extremely high or low point, then researchers should check the outliers to determine whether the data are affected by one-time events.

(IV) Construct an index. Through GDFM, a coincident index is constructed to comprehensively represent a traveler’s interests in a tourist city. This index, which is a combination of the lead and lag information of all search trends data series, reflects the common components of these search trends data.

(V) Predict tourist volumes. Several econometric models are constructed and evaluated. In this research, as a benchmark model, a simple autoregressive moving average model assumes that the current tourist volumes are influenced by past patterns, and thus it is also considered as a benchmark model in the existing literature (Song & Li, 2008). Furthermore, we conduct another benchmark econometric model that obtains the index using PCA. If the GDFM performs the best among the three models, then we can argue that our method is superior in forecasting tourist volumes.

(VI) Evaluate the forecasting accuracy. To verify the forecasting accuracy of different econometric models, we adopt three criteria: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The model with the lowest values on the three criteria is the best forecasting model.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}.
\]
4 Empirical Study

To verify our proposed research framework, we conducted an empirical study on the forecasting of monthly Beijing tourist volumes. Beijing, the capital of China, is one of the most well-known destinations attracting a large volume of domestic and international tourists. In this empirical study, tourist volumes are predicted with Baidu search trends data based on the framework, and the forecasting accuracies of the different models are evaluated. Section 4.1 describes all data sources and data correlations. Section 4.2 presents the empirical results.

4.1 Data

4.1.1 Beijing tourist volumes

Monthly Beijing tourist volumes are collected from the Beijing Tourism Association (BJTA, 2015). The data series, noted as vis, ranges from January 2011 to July 2015. Figure 1 shows the domestic tourist series in Beijing that presents cyclical fluctuations. For convenient modeling, the logs of tourist volumes are computed as logvis.
4.1.2 Baidu search trends data

We take the following steps to select keywords and collect multiple Baidu search trends data to represent travelers’ interests and search behavior.

1. Six major categories are selected in the aspects of visiting Beijing, namely, dining, lodging, traffic, tour, shopping, and recreation. These six categories not only reflect the travelers’ demand but also represent relevant industries on the supply side.

2. Several initial keywords in each category are defined using domain knowledge as the seed keywords. These seeds are used in generating more keywords in the following step.

3. The keyword list is extended using the frequently appearing terms on the search engine interface; around 200 keywords are obtained. The purpose of this step is to maximize the possible keyword pool to represent all aspects of tourists’ interests on tourism cities.

4. The abovementioned keywords are checked manually on the Baidu search engine to ensure the availability of the search trends data series for download. Baidu does not provide search trends data if the volumes of certain keywords are extremely low. In this process, some keywords with unavailable volume data are eliminated.

5. The volumes for search trends data series are obtained. These data series represent weekly frequencies starting from January 1, 2011 to July 25, 2015, as listed in Table 1.

4.1.3 Correlation analysis

Pearson correlations among tourist volumes and all search trends data are computed. Table 1 shows the contemporaneous correlation coefficients between tourist volumes and search trends data. Most search trends data are positively correlated with tourist volumes. However, some search trends data are poorly correlated (e.g., Beijing bar and Beijing recreation places). In existing literature such as Yang, Pan, Evans, and Lv, (2015), poorly correlated series below a certain threshold are removed. Different from their approach, our method keeps the search trends data series in our models to prevent information loss.
### 4.2 Empirical results

In this section, we construct GDFM-based and PCA-based indices using their respective models. We then conduct co-integration and Granger causality tests between tourist volumes and the indices. Using these two indices, we construct two forecasting models and evaluate the out-of-sample forecast accuracy of each.

#### 4.2.1 GDFM-based index and PCA-based index

We first construct the GDFM-based index. The figure of variance contribution rate (shown in the appendix) indicates that four factors can explain most of the variance. Accordingly, we set the number of dynamic factors to 4 and the number of lags of the factors to 5, following the Akaike Information Criterion (AIC) (Forni et al., 2000).
Thereafter, we use Equations 3.1–3.2 to generate the common components (noted as $C_{i,t}$) in the search query data. The common components are described as follows:

$$C_{i,t} = b_{i1}(L^5)f_{i1} + b_{i2}(L^5)f_{i2} + b_{i3}(L^5)f_{i3} + b_{i4}(L^5)f_{i4}.$$  

Eq. 4.1

$$gdfm_i = \frac{\sum_{t=1}^{n} (C_{i,t})}{n}.$$  

Eq. 4.2

In the formulas, $C_{i,t}$ is the coincident index (Forni et al., 2000). To construct the GDFM-based index, we require a new index that can represent all the common components in the search query data. Different from the two methods reviewed in Section 2.3, this new index has two advantages. The GDFM-based index keeps all the original information of search queries. It also considers the lead and lag structures of the common components of search queries to depict their dynamic correlations, as shown in Equation 4.1. The estimation procedure is conducted in MatLab.

For the PCA-based index, we require factors from the search query data. Let $\{x_{i,t}, i = 1, 2, \ldots, n, t = 1, 2, \ldots, T\}$ be the set of search query data. The extracted factors are noted as the following:

$$pca_t = a_{t1}x_{t1} + a_{t2}x_{t2} + \cdots + a_{tp}x_{tn}.$$  

Eq. 4.3

The equation suggests that the new factor is the linear combination of the original data. $p$ is the number of factors, and it should be lower than the sample size $n$. As shown in Table A.1, the first factor explains roughly 30% of the variability, which is the largest proportion of the total factors. Accordingly, we construct the PCA-based index by using the first factor.

### 4.2.2 Co-integration and Granger causality tests

Table 2 shows the stability and the Johansen system co-integration tests among the GDFM-based index, PCA-based index, and tourist volumes. These three data series are stable when validated with the Augmented Dickey–Fuller test. The co-integration results indicate that the GDFM-based index and Beijing tourist volumes are co-integrated. Similarly, the PCA-based index and Beijing tourist volumes are co-integrated. Furthermore, a long-term relationship exists between the search trends data and tourist volumes. Therefore, the findings suggest the feasibility of adopting search trends data series in the econometric models.

The purpose of the Granger causality tests is to verify whether the search trends index is predictive of tourist volumes. As shown in Table 3, the GDFM-based index and the PCA-based index are Granger causal of tourist volumes. This finding indicates that search trends data lead the actual Beijing tourist volumes.
Table 2. Co-integration test results

<table>
<thead>
<tr>
<th>Augmented Dickey–Fuller tests</th>
<th>t statistics</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>logvis</td>
<td>-3.7963</td>
<td>0.0034</td>
</tr>
<tr>
<td>GDFM index</td>
<td>-3.6844</td>
<td>0.0049</td>
</tr>
<tr>
<td>PCA index</td>
<td>-3.3980</td>
<td>0.0120</td>
</tr>
</tbody>
</table>

Cointegration between logvis and GDFM-based index

<table>
<thead>
<tr>
<th></th>
<th>Eigenvalue</th>
<th>Trace statistic</th>
<th>Critical value</th>
<th>Prob**</th>
</tr>
</thead>
<tbody>
<tr>
<td>None *</td>
<td>0.06</td>
<td>28.29</td>
<td>15.49</td>
<td>0.00</td>
</tr>
<tr>
<td>At most 1 *</td>
<td>0.06</td>
<td>13.42</td>
<td>3.84</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Cointegration between logvis and PCA-based index

<table>
<thead>
<tr>
<th></th>
<th>Eigenvalue</th>
<th>Trace statistic</th>
<th>Critical value</th>
<th>Prob**</th>
</tr>
</thead>
<tbody>
<tr>
<td>None *</td>
<td>0.07</td>
<td>18.04</td>
<td>15.49</td>
<td>0.02</td>
</tr>
<tr>
<td>At most 1 *</td>
<td>0.00</td>
<td>0.84</td>
<td>3.84</td>
<td>0.36</td>
</tr>
</tbody>
</table>

* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug Michelis (1999) p-values

Table 3. Granger causality tests between the constructed index and tourist volumes

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>F-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDFM-based index does not Granger Cause logvis</td>
<td>34.08</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>logvis does not Granger Cause GDFM-based index</td>
<td>0.40</td>
<td>0.53</td>
</tr>
<tr>
<td>PCA-based index does not Granger Cause logvis</td>
<td>26.33</td>
<td>0.00 ***</td>
</tr>
<tr>
<td>logvis does not Granger Cause PCA-based index</td>
<td>0.01</td>
<td>0.94</td>
</tr>
</tbody>
</table>

*** indicates the significance level at 1%

4.2.3 Econometric modeling

Three econometric models are conducted to examine the predictive power of the search trends data index. In accordance with the existing literature, Model 1, a simple autoregressive moving averaging model, is selected as benchmark model 1. This model assumes that the current values of time series are affected by its past patterns.

In addition to Model 1, Model 2 is constructed as the other benchmark model that incorporates the PCA-based index as the independent variable. Model 3 uses the GDFM-based index to predict Beijing tourist volumes. Both lags of the GDFM- and PCA-based indices are set to 5 according to the AIC criterion. We equalize the number of independent variables in Model 2 and Model 3 to evaluate them on the same condition.

Model 1: \( \text{Logvis}_t = \alpha_1 + \alpha_2 \text{Logvis}_{t-1} + \alpha_3 \varepsilon_{t-1} + \varepsilon_t \).

This model serves as benchmark model 1 and considers the first lag of the dependent variable.

Model 2: \( \text{Logvis}_t = \sum_{i=1}^{5} \beta_{1,i} \text{pca}_{t-i} + \beta_0 + \beta_1 \text{Logvis}_{t-1} + \beta_2 \varepsilon_{t-1} + \varepsilon_t \).
In Model 2, the independent variables in Model 1 are retained, and the lags of \( pca_t \), which is computed using Eq. 4.3, are added.

Model 3: 
\[
\text{Logvis}_t = \sum_{i=1}^{5} \gamma_{i,t} \text{gdfm}_{t-i} + \gamma_6 \text{Logvis}_{t-1} + \gamma_7 e_{t-1} + \epsilon_t.
\]

Model 3 incorporates the GDFM-based index obtained from Eq. 4.2. In accordance with Model 2, the lag of this index is set to 5.

In the abovementioned models, \( \text{Logvis}_t \) indicates the tourist arrivals; \( \alpha_1, \alpha_2, \alpha_3 \) indicate the coefficients of Model 1; and \( \beta_1, \beta_2, \ldots, \beta_6 \) and \( \gamma_1, \gamma_2, \ldots, \gamma_6 \) represent the estimated coefficients of Models 2 and 3, respectively.

Table 4 presents the estimated coefficients and the key measurements of these three models. Adjusted \( R^2 \) describes the fitness of the econometric models, and it is a more valid measurement of fitness than \( R^2 \). AIC and SC are information criteria, which also characterize the performance of the models. The model with the lowest AIC and SC has the best performance. The bolded values in Table 4 indicate the model with the largest goodness-of-fit and the lowest information criteria.

As shown in Table 4, Model 3 with the GDFM-based index has the highest adjusted \( R^2 \) and the lowest AIC and SC criteria. Mode 3 has an adjusted \( R^2 \) improved by 1.57% and 0.11% compared with the benchmark and the model with the PCA-based index, respectively.

In terms of the estimated coefficients, all the independent variables are positively correlated with tourist volumes. In Model 3, a 1% increase in the lags of the GDFM-based index leads to approximately 0.1% increase in the variance of the dependent variable, with other variables unchanged. The coefficients of the GDFM-based index in Model 3 are significantly higher than those in Model 2, thus suggesting that the GDFM-based index contains greater explanatory power than the PCA-based index.

<table>
<thead>
<tr>
<th>Model 1 Variables</th>
<th>Coefficients</th>
<th>Model 2 Variables</th>
<th>Coefficients</th>
<th>Model 3 Variables</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_1 )</td>
<td>0.0027 ***</td>
<td>( \gamma_1 )</td>
<td>0.0206 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>0.0024 ***</td>
<td>( \gamma_2 )</td>
<td>0.0189 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>0.0023 ***</td>
<td>( \gamma_3 )</td>
<td>0.0184 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>0.0024 ***</td>
<td>( \gamma_4 )</td>
<td>0.0191 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>0.0019 ***</td>
<td>( \gamma_5 )</td>
<td>0.0149 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_1 )</td>
<td>7.6943 ***</td>
<td>( \beta_6 )</td>
<td>5.9833 ***</td>
<td>( \gamma_6 )</td>
<td>5.6913 ***</td>
</tr>
<tr>
<td>( \alpha_2 )</td>
<td>0.8829 ***</td>
<td>( \beta_7 )</td>
<td>0.9466 ***</td>
<td>( \gamma_7 )</td>
<td>0.9275 ***</td>
</tr>
</tbody>
</table>

Table 4. Estimation of econometric models
4.2.4 Out-of-sample forecasting evaluation

Static and dynamic checks on the out-of-sample forecasting power are conducted to evaluate the forecasting accuracy of the models. First, we conduct static one-week and four-week out-of-sample forecasts, in which the sample periods from January 2011 to July 18 2015 are estimated, and forecast the value on July 25, 2015. The major purpose of the four-week out-of-sample forecasts is to examine whether the model has a relatively long-term forecasting ability.

Second, for the robustness of the evaluation, dynamic rolling window forecasts are also conducted to assess the forecasting accuracy of the econometric models. We set the window length, which is the estimated sample periods, to 180 weeks and run one-week and four-week forecasts. In each forecast, the estimated samples and forecasting points are considered as dynamic to test the robustness of the model.

Table 5 shows the RMSE, MAE, and MAPE of these econometric models. To compare the performance of Model 3 with those of the other two models, we compute the improvement rate of the forecasting accuracy from Model 1–2 to Model 3. As shown in Table 5, Model 3 outperforms the benchmark model 1 and Model 2 in all the tests. In the static forecasts, the model with the GDFM-based index has a forecast accuracy improved by an average of at least 50% compared with the model with the PCA-based index. In the dynamic forecasts, Model 3 still has a forecast accuracy improved by 28% and 2.5% compared with the benchmark Model 1 and Model 2, respectively. In addition, the short-term forecasting ability is superior to the long-term forecasts. Overall, the GDFM-based index significantly improves the accuracy of tourist volume forecasting.

<table>
<thead>
<tr>
<th>Static out-of-sample forecasting</th>
<th>Forecasting accuracy improvement rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Forecasting step = 1</strong> (Forecasting period: July 25, 2015)</td>
<td><strong>Forecasting step = 4</strong> (Forecasting periods: July 4–25, 2015)</td>
</tr>
<tr>
<td><strong>Model 1</strong></td>
<td><strong>Model 2</strong></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.1084</td>
</tr>
<tr>
<td>MAE</td>
<td>0.1084</td>
</tr>
<tr>
<td>MAPE</td>
<td>1.3164</td>
</tr>
<tr>
<td><strong>Dynamically rolling forecast, window length = 180 weeks</strong></td>
<td></td>
</tr>
</tbody>
</table>

*** indicates the significance level of 1%
## 5 Conclusion and Implications

This research proposed a new forecasting framework with search trends data and applied it to the prediction of Beijing tourist volumes. First, we introduce a GDFM that uses the common components of search trends data to construct a more comprehensive index. Second, we compare this new index with a time series model and the PCA-based index model commonly used in existing studies. We evaluate the performances of the econometric models with different indices using static and dynamic tests.

The empirical study indicates that our framework has a more favorable performance than other econometric models. First, a significant co-integration relationship exists between the index and Beijing tourist volumes. Second, Granger causality tests suggest that search trends data lead the actual tourist volumes. Third, we demonstrate that the econometric model with the new index has the best forecasting accuracy in the one-week and four-week forecasts. We also conduct the rolling window forecasts for the robustness check. The empirical results validated our framework, which offers a suitable solution for better manipulating large-scale search trends data.

Our research has theoretical implications. We propose a theoretical framework for search trends data index aggregation based on GDFM. A large dataset of search trends data usually has complicated correlations. Different from the aggregation in the study of Yang, Pan, Evans, and Ly (2015), the aggregation of this index is computed through the lead–lag orders among these search trends data. We compute common components in all search trends data and construct an integrated index accordingly. The method can better manipulate large datasets of search trends data in a simple and flexible manner.

Furthermore, timely and accurate tourist forecasts are crucial for policy makers and business managers. This study indicates that the search trends data index provides more accurate forecasts of tourist volumes than other indices. Policy makers should monitor this aggregated index to better capture the dynamics of tourist volumes. In addition, our empirical study demonstrates that this new index has the best forecasting performance in short-term forecasts. Accurate forecasts can offer useful support to businesses in making the most strategic decisions during peak or off-peak tourism seasons.
This research has several limitations, and some of them can be investigated for future research. Although we expect that the new forecasting framework can be extended to other domains, we still need further rigorous experiments to examine whether the new methodology can predict other important indicators in tourism and hospitality such as tourist volumes in other cities, hotel sales, and flights bookings. We believe that a future study is meaningful when it addresses the importance of index aggregation in various domains. We hope that this research will encourage future studies to better process large search trends datasets for more accurate forecasts. Another limitation of this study is that it mainly focuses on the econometric models with search trends data. In fact, many effective machine learning approaches are used to model large datasets. Therefore, future studies should investigate whether our method is useful in nonlinear models.

References


Appendix A

Table A.1: Results of PCA

<table>
<thead>
<tr>
<th>Number</th>
<th>Value</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative Value</th>
<th>Cumulative Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.7479</td>
<td>1.3265</td>
<td>0.3055</td>
<td>13.7479</td>
<td>0.3055</td>
</tr>
<tr>
<td>2</td>
<td>12.4214</td>
<td>8.9336</td>
<td>0.2760</td>
<td>26.1692</td>
<td>0.5815</td>
</tr>
<tr>
<td>3</td>
<td>3.4877</td>
<td>0.5389</td>
<td>0.0775</td>
<td>29.6570</td>
<td>0.6590</td>
</tr>
<tr>
<td>4</td>
<td>2.9488</td>
<td>1.4649</td>
<td>0.0655</td>
<td>32.6058</td>
<td>0.7246</td>
</tr>
<tr>
<td>5</td>
<td>1.4839</td>
<td>0.3688</td>
<td>0.0330</td>
<td>34.0896</td>
<td>0.7575</td>
</tr>
<tr>
<td>6</td>
<td>1.1151</td>
<td>0.0711</td>
<td>0.0248</td>
<td>35.2048</td>
<td>0.7823</td>
</tr>
<tr>
<td>7</td>
<td>1.0440</td>
<td>0.0675</td>
<td>0.0232</td>
<td>36.2488</td>
<td>0.8055</td>
</tr>
<tr>
<td>8</td>
<td>0.976454</td>
<td>0.121597</td>
<td>0.0217</td>
<td>37.22521</td>
<td>0.8272</td>
</tr>
<tr>
<td>9</td>
<td>0.854856</td>
<td>0.14662</td>
<td>0.019</td>
<td>38.08007</td>
<td>0.8462</td>
</tr>
<tr>
<td>10</td>
<td>0.708236</td>
<td>0.085313</td>
<td>0.0157</td>
<td>38.7883</td>
<td>0.862</td>
</tr>
</tbody>
</table>

Fig. A.1. Variance contribution rate

Fig. A.2. GDFM-based index and PCA-based index
Author Biography

Xin Li, Ph.D., is a researcher at e-Tourism Research Center, Institute of Tourism, Beijing Union University. Dr. Li’s research interests are big data analytics, econometric modeling, data mining and forecasting. Dr. Li focuses on understanding tourism activities by combing user-generated contents with econometric and machine learning techniques. She also participated in many research projects on monitoring, forecasting, and earlywarning of economy and industries in China.

Bing Pan, Ph.D., is an Associate Professor in the Department of Hospitality and Tourism Management and Head of Research in the Office of Tourism Analysis within the School of Business at the College of Charleston, USA. His research publications include using online data to understand, predict, monitor, and forecast tourism economic activities, tourist online behavior, social media, search engine marketing, and research methodologies. Dr. Pan has consulted with the Charleston Area Convention and Visitors Bureau for nine years.
Rob Law, Ph.D., is a Professor at the School of Hotel and Tourism Management, the Hong Kong Polytechnic University. His research interests are information management, modelling and forecasting, artificial intelligence and technology applications. He has received many research related awards and honors, as well as millions of USD external and internal research grants. Prof. Law serves in different roles for 140+ research journals, and is a chair/committee member of more than 130 international conferences.

Xiankai Huang, Ph.D., is a Professor at Beijing Open University. Prof. Huang’s research interests are statistical modeling and tourism forecasting. Prof. Huang has hosted many national programs such as National Key Technology Research and Development Program of the Ministry of Science and Technology of China, with over millions of research grants.