Tourism Management 70 (2019) 1–10

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman

# Forecasting tourist arrivals with machine learning and internet search index

Shaolong Sun[a,b,c], Yunjie Wei[a,d], Kwok-Leung Tsui[c], Shouyang Wang[a,b,d,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China
[b] School of Economics and Management, University of Chinese Academy of Sciences, Beijing, 100190, China
[c] Department of Systems Engineering and Engineering Management, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[d] Center for Forecasting Science, Chinese Academy of Sciences, Beijing, 100190, China

ARTICLE INFO

*Keywords:*
Tourism demand forecasting
Kernel extreme learning machine
Search query data
Big data analytics
Composite search index

ABSTRACT

Previous studies have shown that online data, such as search engine queries, is a new source of data that can be used to forecast tourism demand. In this study, we propose a forecasting framework that uses machine learning and internet search indexes to forecast tourist arrivals for popular destinations in China and compared its forecasting performance to the search results generated by Google and Baidu, respectively. This study verifies the Granger causality and co-integration relationship between internet search index and tourist arrivals of Beijing. Our experimental results suggest that compared with benchmark models, the proposed kernel extreme learning machine (KELM) models, which integrate tourist volume series with Baidu Index and Google Index, can improve the forecasting performance significantly in terms of both forecasting accuracy and robustness analysis.

## 1. Introduction

All over the world, the tourism industry contributes significantly to economic growth (Gunter & Onder, 2015; Song, Li, Witt, & Athanasopoulos, 2011). According to the China National Tourism Administration, in 2016 the tourism income of China reached 4.69 trillion RMB, increasing by 13.6% compared to the previous year, and accounted for 6.3% of China's GDP. Thus, forecasting tourist volume is becoming increasingly important for predicting future economic development. Tourism demand forecasting may provide basic information for subsequent planning and policy making (Chu, 2008; Witt & Song, 2002). Methods used in tourism modeling and forecasting fall into four groups: time series models, econometrics models, artificial intelligence techniques and qualitative methods (Goh & Law, 2011; Song & Li, 2008). In addition to simple tourist data announced by the State Statistics Bureau, Internet search queries, which reflect the behavior and intentions of tourists, have increasingly been used in tourism forecasting models (Croce, 2017; Goodwin, 2008). However, the search index has created big opportunities in the modeling process of tourism forecasting (Li, Pan, Raw & Huang, 2017).

Internet search data has been applied to many aspects, such as hotel registrations (Pan & Yang, 2017; Rivera, 2016), tourist numbers (Bangwayo-Skeete & Skeete, 2015; Yang, Pan, Evans, & Lv, 2015), economic indicators (Choi & Varian, 2012), unemployment rates (Askitas & Zimmermann, 2009), private consumption (Vosen & Schmidt, 2011), and stock returns (Zhu & Bao, 2014). When introducing the Baidu Index or Google Index into forecasting models, keywords and the composition of indexes must be selected carefully. Keywords can be selected according to the correlation coefficient, the tendency chart or the crowd-squared method (Brynjolfsson, Geva, & Reichman, 2016). Additionally, the composition of indexes can be achieved by the HE-TDC method (Peng, Liu, Wang, & Gu, 2017) or the principal component analysis (PCA). Obviously, efforts should be made to avoid problems related to multi-collinearity and over-fitting to the greatest extent possible.

In this study, we proposed a new framework integrating machine learning and Internet search index to forecast tourist volume. The forecasting power of the framework is attributable to two features: first, relevant Internet search queries greatly contribute to the goodness of fit; second, Kernel-based extreme learning machines have short computing time and good generalization ability. However, as far as we know, few studies have adopted extreme learning machine to forecast tourism demand. The proposed framework is utilized to forecasting Beijing tourist arrivals. Relevant Internet search keywords cover the various aspects of tourism including dining, lodging, recreation, shopping, tour and traffic. Different from previous studies, this paper considers both Baidu Index and Google Index, which reflect the current situation of domestic tourists and foreign travelers. The experimental

* Corresponding author. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhongguancun East Road, #55, Haidian District, Beijing, 100190. China.
*E-mail addresses:* sunshl@amss.ac.cn (S. Sun), weiyunjie@amss.ac.cn (Y. Wei), kltsui@cityu.edu.hk (K.-L. Tsui), sywang@amss.ac.cn (S. Wang).

https://doi.org/10.1016/j.tourman.2018.07.010
Received 8 February 2018; Received in revised form 29 May 2018; Accepted 13 July 2018

**Table 1**
An overview of selected tourism forecasting studies.

| References | Region focused | Research objects | Data frequency | Methodologies | Performance measure | Variables |
|---|---|---|---|---|---|---|
| Athanasopoulos and Hyndman (2008) | Australian | Inbound tourism | Quarterly | SSME, ES | RMSE, ME, MAE, MAPE | Tourist arrivals and economic variables |
| Bangwayo-Skeete & Skeete, 2015 | Caribbean | Tourism demand | Monthly | AR-MIDAS | MAPE, RMSE, DM | Tourist arrivals, Google trend data |
| Chen, Lai, and Yeh (2012) | Taiwan | Inbound tourism | Monthly | EMD, BPNN | MAD, MAPE, RMSE | Tourist arrivals |
| Chu (2008) | Asian-Pacific | Tourism demand | Monthly, quarterly | ARAR model | MAPE, RMSE | Tourist arrivals |
| Fildes, Wei, and Ismail (2011) | UK | Air travel demand | Annual | ADLM, TVP, VAR | MAE, RMSE | Air passengers and economic variable |
| Gunter and Onder (2015) | Paris | Inbound tourism | Monthly | EC-ADLM, VAR, Bayesian VAR, TVP, ARMA, ETS | RMSE, MAE | Tourist arrivals in hotels and economic variable |
| Jungmittag (2016) | German | Travel demand | Monthly | Combination forecasts, SARIMA | MAE, RMSE, MAPE | Air passengers |
| Li, Pan, Law, and Huang (2017) | Beijing | Tourism demand | Monthly | GDFM, PCA | MAE, MAPE | Tourist arrivals and Baidu index |
| Liang (2014) | Taiwan | Inbound tourism | Monthly | SARIMA, GARCH | MAD, RMSE, MAPE | Tourist arrivals |
| Pan and Yang (2017) | Charleston county | Hotel demand | Weekly | ARIMAX | MAPE, RMSE | Hotel occupancy, search engine queries, website traffic, weather information |
| Du Preez and Witt (2003) | Seychelles | Inbound tourism | Monthly | ARIMA, SSM | MAE, RMSE, MAPE | Tourist arrivals and economic variables |
| Rivera (2016) | Puerto Rico | Hotel nonresident registrations | Monthly | DLM | MAE, MAPE, RMSE | Google trend data, NHNR |
| Shahrabi, Hadavandi, and Asadi (2013) | Japan | Inbound tourism | Monthly | MGFFS | RMSE, MAPE | Tourist arrivals |
| Song et al. (2011) | Hong Kong | Inbound tourism | Monthly | STSM, TVP | MAPE, RMSE | Tourist arrivals and economic variables |
| Wong, Song, Witt, and Wu (2007) | Hong Kong | Inbound tourism | Quarterly | ARIMA, ADLM, ECM, VAR, combining forecast | MAPE, RMSE, MAE | Tourist arrivals and economic variables |
| Wu and Cao (2016) | Mainland China | Inbound tourism | Monthly | SVR, FOA, SIA | MAPE, RMSE, R | Inbound tourist flow |

results illustrate that the proposed forecasting framework is significantly superior to the traditional time series models and some other machine learning models. Meanwhile, the forecasting power of the models with Baidu Index and Google Index is stronger than that without one index or both indexes, which may provide solid evidence that Internet search queries are of great significance to tourism demand forecasting.

The remainder of this paper is organized as follows: Literature review is provided in Section 2. Kernel extreme learning machine is introduced in Section 3. Forecasting framework is shown in Section 4. The empirical study is given in Section 5. Finally, Section 6 offers concluding work and implications for further research.

## 2. Literature review

This section reviews relevant literature about tourist arrivals forecasting and tourism forecasting with search engine query data. A list of these literature is provided in Table 1.

NOTE: state space models with exogenous variables (SSME) model; exponential smoothing (ES) model; Autoregressive Mixed-Data Sampling (AR-MIDAS) models; empirical mode decomposition (EMD); back propagation neural network (BPNN); autoregressive distributed lag model (ADLM); time-varying parameter (TVP); vector autoregressive (VAR) model; error correction autoregressive distributed lag model (EC-ADLM); Bayesian vector autoregressive (BVAR); autoregressive moving averaging (ARMA); seasonal autoregressive integrated moving average (SARIMA); generalized dynamic factor model (GDFM); principal component analysis (PCA); generalized autoregressive conditional heteroskedasticity (GARCH); autoregressive integrated moving average with exogenous variables (ARIMAX); state space models (SSM); Dynamic linear model (DLM); modular genetic-fuzzy forecasting system (MGFFS); structural time series model (STSM); support vector regression (SVR); fruit fly optimization algorithm (FOA); seasonal index adjustment (SIA); root mean square error (RMSE); mean error (ME); mean absolute error (MAE); mean absolute percentage error (MAPE); Diebold-Mariano (DM) statistic; mean absolute deviation (MAD); the number of hotel nonresident registrations (NHNR).

### 2.1. Tourist volume forecasting

Autoregressive integrated moving average (ARIMA) is the most widely used time series forecasting model. This model has also been widely applied to tourism forecasting and performed well (Athanasopoulos, Hyndman, Song, & Wu, 2011; Brida & Risso, 2011; Chang & Liao, 2010; Chen et al., 2012; Du Preez & Witt, 2003; Jungmittag, 2016; Li & Sheng, 2016; Liang, 2014; Lim & McAleer, 2002; Shahrabi et al., 2013). However, ARIMA models do not always outperform others. These models perform well in the traditional econometric models, but they are sometimes inferior to intelligence methods. Song and Witt (2000) used a variety of techniques to predict tourism demand for a specified region and found that a neural network approach outperformed ARIMA model. Similarly, previous research on tourist arrivals in China demonstrated that support vector regression (SVR) outperformed back propagation neural network (BPNN) and ARIMA models.

Exponential smoothing (ES) has been widely used in tourism forecasting and many scholars use the model as a benchmark (Fildes et al., 2011; Park, Rilett, & Han, 1999; Witt & Witt, 1995). Other time series models include but are not limited to state space models (Athanasopoulos & Hyndman, 2008; Beneki, Eeckels, & Leon, 2012; Du Preez & Witt, 2003), error correction models (ECM) (Lee, 2011; Shen, Li, & Song, 2009; Vanegas, 2013; Wong et al., 2007), and generalized autoregressive conditional heteroskedastic (GARCH) models (Chan, Lim, & Mcaleer, 2005; Liang, 2014).

In recent years, artificial intelligence (AI) techniques have emerged in the tourism study, such as fuzzy logic theory, artificial neural