Tourism Management 83 (2021) 104245

Contents lists available at ScienceDirect

Tourism Management

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

Progress in Tourism Management

Review of tourism forecasting research with internet data

Xin Li [a,*], Rob Law [b], Gang Xie [c], Shouyang Wang [c]

[a] School of Economics and Management, University of Science and Technology Beijing, 100083, China
[b] School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Hong Kong SAR, China
[c] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China

ARTICLE INFO

Keywords:
Internet data
Tourism forecasting
Search engine
Social media
Systematic review

ABSTRACT

Internet techniques significantly influence the tourism industry and Internet data have been used widely used in tourism and hospitality research. However, reviews on the recent development of Internet data in tourism forecasting remain limited. This work reviews articles on tourism forecasting research with Internet data published in academic journals from 2012 to 2019. Then, the findings ae synthesized based on the following Internet data classifications: search engine, web traffic, social media, and multiple sources. Results show that among such classifications, search engine data are most widely incorporated into tourism forecasting. Time series and econometric forecasting models remain dominant, whereas artificial intelligence methods are still developing. For unstructured social media and multi-source data, methodological advancements in text mining, sentiment analysis, and social network analysis are required to transform data into time series for forecasting. Combined Internet data and forecasting models will help in improving forecasting accuracy further in future research.

1. Introduction

The increasing growth of Internet applications, such as social media platforms, has generated overwhelming data in different formats, including structured and unstructured (Song & Liu, 2017). The second generation of the Internet began after 2004 and Web 2.0-based applications provide efficient platforms that generate large volumes of data (Leung, Law, Van Hoof, & Buhalis, 2013). Such Internet data are considered important big data that can be obtained from different sources, including the Internet, sensors, transactions, and the Internet of things (Hashem et al., 2015). With the continuous increase in the number of Internet applications, abundant sources can generate Internet data, including search engines, blogs, microblogs (e.g., Twitter), social networks (e.g., Facebook), photo sharing (e.g., Flickr), and video sharing (e.g., YouTube) (Berthon, Pitt, Plangger, & Shapiro, 2012; Li, Xu, Tang, Wang, & Li, 2018). Internet applications provide tourists with online channels from which to retrieve information, express opinions, and book trips. Accordingly, various real-time Internet data reflecting tourist behaviors are produced in different formats during this process. As a type of big data that can significantly influence the tourism and hospitality industry, new and ample Internet data are considered important supplements to traditional data in different industries and academia (Li & Law, 2020; Song, Qiu, & Park, 2019). Existing literature indicates that analysis and modeling enable the use of data to understand consumer behaviors and support marketing or management decision-making for tourism industries (Li, Xu, et al., 2018).

Normally, Internet data are generated from the web and social media sources in virtual networks, such as search engines, Wikipedia, microblogs, Facebook, and Twitter (Hashem et al., 2015; Song & Liu, 2017). For example, search volumes of specific keywords in search engines represent Internet data that reflect users' attention and interests (Li, Pan, Law, & Huang, 2017). The posts, reposts, user reviews, and photos are Internet data that represent user behavior. Li, Xu, et al. (2018) indicated that online data constituted the largest shares of big data (55%) used in tourism research. Song et al. (2019) suggested that Internet data are viewed as a new driving factor in tourism-forecasting research. In recent years, researchers have widely incorporated such Internet data into forecasting practices as explanatory variables in the tourism and hospitality industry in recent years (Song et al., 2019). Given the increasing importance and complexity of tourism forecasting, researchers contribute to improving forecasting accuracy further by proposing advanced methods. Song et al. (2019) summarized the methodological development of forecasting models from 1968 to 2018 and showed that forecasting methods continue to evolve. However, the development and evolution of tourism forecasting with different types of Internet data remain unclear. In particular, methods for analyzing and modeling

* Corresponding author. 30 Xueyuan Road, Haidian District, Beijing, 100083, China.
E-mail addresses: drxinli@ustb.edu.cn (X. Li), rob.law@polyu.edu.hk (R. Law), gxie@amss.ac.cn (G. Xie), sywang@amss.ac.cn (S. Wang).

https://doi.org/10.1016/j.tourman.2020.104245
Received 16 January 2020; Received in revised form 25 August 2020; Accepted 12 October 2020
Available online 21 October 2020

structured and unstructured Internet data may differ and the forecasting accuracy of each kind of data varies considerably.

The primary goal of this study is to conduct a comprehensive review of the research state of Internet data used in tourism forecasting and to trace the methodological development of four types of Internet data (i.e., search engine, web traffic, social media, and multi-source data) in the field of tourism. The first article on tourism forecasting with Internet data was published in 2012. Hence, all relevant articles in academic journals from the tourism and hospitality literature for the period from 2012 to 2019 were collected. This study aims to investigate how forecasting models, including time series, econometrics, artificial intelligence, and hybrid methods are adopted in addressing different Internet data. It attempts to determine the differences between the four types of Internet data for tourism forecasting and explore how tourism-forecasting accuracy can be improved based on Internet data. It also provides implications for the use of Internet data in tourism forecasting.

The rest of the paper is organized as follows. Section 2 illustrates the detailed process of key literature selection and describes the categories of Internet data. Section 3 presents the review findings from the perspectives of the search engine, web traffic, social media, and multi-source data with a focus on modeling techniques. Section 4 discusses the findings and implications. The last section provides the conclusion and offers potential future directions and current limitations.

2. Literature selection

In October 2019, the relevant literature on tourism forecasting with Internet-generated data was identified and collected from large databases, including Web of Science, Science Direct, EBSCOHost, and Google Scholar. Only full-length articles in the English language in academic journals were included, and reviews, conference articles, editorial, abstracts, and letters were excluded. We also traced the references in the published articles to ensure all relevant literature was considered. We used keywords, including *Internet data, web data, web traffic, search data, social media, social network, tourism forecasting/prediction*, *hotel forecasting/predicting, tourist arrivals forecasting/predicting,* and *hotel occupancy forecasting/predicting* to provide a broad overview of research trends in tourism forecasting with Internet data to search for the most relevant articles. All retrieved articles were assessed further by an expert panel including two academic and research staff in tourism forecasting fields to ensure the accuracy and objectivity of the review findings. Finally, 47 published studies were included in the final analysis.

Pioneering research articles related to tourism forecasting using Internet data started in 2012, when Pan, Wu, and Song (2012) and Choi and Varian (2012) collected Internet data from Google Trends to predict hotel rooms in Charleston, US and tourist arrivals to Hong Kong, China. The numbers of such relevant studies were one, one, five, and two from 2013 to 2016. These studies experienced significant growth from 2017 to 2019, when the numbers increased to 9, 10, and 17, respectively, until the end of October 2019. Fig. 1 shows that such numbers indicate Internet data have attracted increasing attention from researchers in the fields of tourism and hospitality forecasting, particularly in recent years.

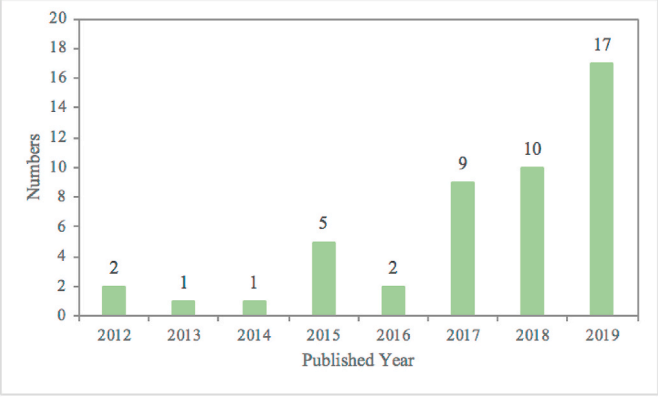

Fig. 1. Number of selected articles from 2012 to 2019.

The selected articles were published in different academic journals in various research fields, including tourism and hospitality, economics, management, business, and computer science. These tourism and hospitality-related journals include *Tourism Management* (11), *Tourism Economics* (5), *Journal of Travel Research* (4), *International Journal of Hospitality Management* (3), *Annals of Tourism Research* (2), *Journal of Travel and Tourism Marketing* (1), *Asia Pacific Journal of Tourism Research* (1), and *Tourism Management Perspectives* (1). Other research fields including economic, business, and information management also published such relevant tourism-forecasting articles in journals, such as *Economic Record* (1), *Applied Economics* (1), *Decision* Support *Systems* (1), *Journal of Environmental Management* (1*), Applied Soft Computing Journal* (1), *Journal of Information Science* (1), *Technological Forecasting and Social Change* (1), and *Information and Management* (1).

Fig. 2 presents the classifications of Internet data based on the data source, format, frequency, and forecasting technique (Song & Liu, 2017). These relevant tourism-forecasting articles are classified into four categories based on the data used (e.g., search engine, web traffic, social media, and multi-source data) and the differences in generated sources of Internet data. Detailed information on the selected studies is summarized in Table 1. The key findings and discussions on the research articles are provided in the following section.

ADE: Adaptive differential evolution algorithm; ARX: Autoregressive with exogenous variables; ARMAX: Autoregressive moving average with exogenous variables; ARIMA: Autoregressive integrated moving average; AR-MIDAS: Autoregressive mixed data-sampling approach; ADL: Autoregressive distributed lag; ANOVA: Analysis of variance; ANN: Artificial neural network; ASA: Aspect-based sentiment analysis; BA-SVR: Bat algorithm–support vector regression; BFVAR: Bayesian factor-augmented vector autoregressive; BM: Bridge model; BR: Boosted regression; BVAR: Bayesian vector autoregressive; CA: Cluster analysis; DBN: Deep belief network; DFM: Dynamic factor model; DLM: Dynamic liner model; DM: Demand model; EEMD: Ensemble empirical mode decomposition; EM: Exponential model; ESN: Echo state network; ETS: Exponential smoothing; FAAR: Factor augmented AR; FABM: Factor Augmented bridge model; FAVAR: Factor-augmented vector autoregressive; FCM: Forecast combination methods; FOA-BP: Fruit fly optimization algorithm-Back Propagation Network; GA: Genetic algorithm; GM: Gravity model; HE-TDC: Hurst exponent–time difference correlation; HW: Holt-Winter; KELM: Kernel extreme learning machine; LR: Linear regression; LSTM: Long-short-term memory; MLR: Multiple linear regression; MSDR: Markov switching dynamic regression; NAR: Nonlinear autoregressive; PPR: Projection pursuit regression; PSO: Particle swarm optimization; QM: Quadratic model; SAREN: Stacked autoencoder with echo-state regression; SARIMAI: Seasonal autoregressive integrated moving average with inventions; SVR: Support vector regression; SEM: Structural equation modeling; SITA: Supplementary information transformation; SLFN: Single layer feedforward neural network; SNAIVE: Seasonal naïve model; TAR: Threshold autoregressive; TVP: Time-varying parameter.

3. Review findings

Given the differences in the types of Internet data, studies on tourism forecasting with Internet data are classified into four streams based on the data they used for modeling. Tourism forecasting research on search engine data comprises the majority of reviewed articles at 53%. Data from search engines are structured time-series data that reflect user attention on certain topics through keywords (Li et al., 2017). Social media and multi-source data comprised nearly 20% of the total articles, respectively. The forecasting articles on web traffic data are relatively limited as compared with other data sources. Studies on each kind of

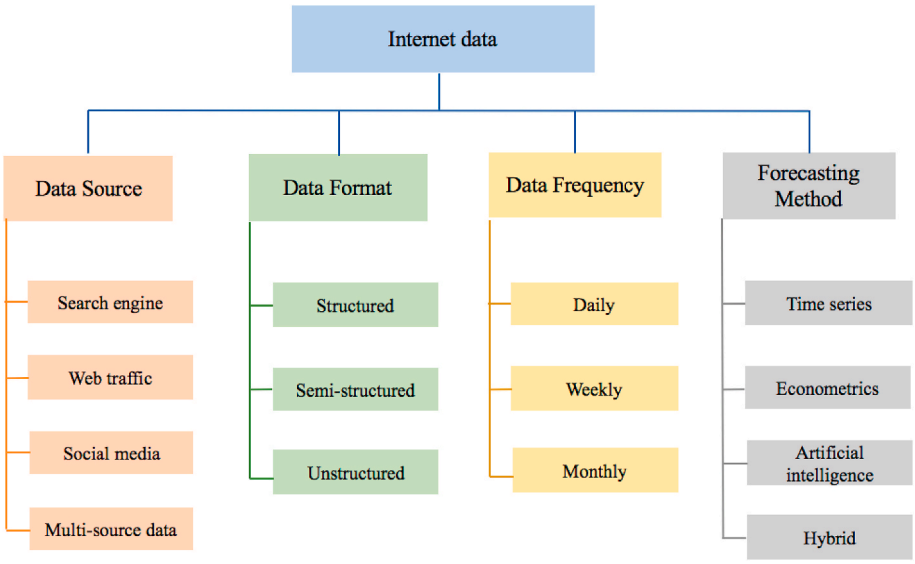

Fig. 2. Classification of Internet data.

data are reviewed and discussed further in the subsequent section. Regarding the discrepancies in the studied contexts, 79% of the reviewed articles focus on forecasting tourist arrivals to travel destinations and 21% are related to hotel demand forecasting, such as hotel occupancy rate.

Fig. 3 highlights the major procedure of tourism forecasting with Internet data, which includes data selection, data extraction, processing and transformation, forecasting, and evaluation. Three steps, which include data selection, data extraction, processing, and transformation to address the Internet data, are consistent with Brynjolfsson, Geva, and Reichman (2016) and described a typical prediction process using online data for prediction. The other steps provide instructions on the construction of forecasting models and the evaluation of forecasting accuracy. The findings of tourism forecasting research with different Internet data are summarized based on the major steps.

3.1. *Forecasting with search engine data*

Search engine data are real-time daily, weekly, and monthly search queries that users enter into search engines, such as Google and Baidu (Choi & Varian, 2012). These structured Internet data are new data sources for tourism forecasting (Song et al., 2019). In the 25 reviewed papers related to forecasting with search engine data, the search query index or volumes were generated from Google Trends and Baidu Index (Yang et al., 2015a). Nine articles (36%) under this category used Baidu data, and 16 articles (64%) adopted Google data for forecasting. Baidu data have been used widely in the forecasting of tourist arrivals to destinations (Li et al., 2017, 2018), scenic areas (Huang et al., 2017; Li et al., 2019), and hotels (Zhang, Pu, & Wang, 2019) in China, including Chinese tourists to other countries, such as Thailand (Tang, 2018). Meanwhile, Google data were applied mainly in forecasting tourism arrivals and hotel occupancy to US, Spain, Italy (Florence, Milan, and Catania), Germany, United Kingdom, South Korea, China, Hong Kong, and Macau from Western countries (Choi & Varian, 2012; Emili, Figini, & Guizzardi, 2019; Law et al., 2019; Park et al., 2017; Rivera, 2016).

Given the differences between predicted contexts, search engine data at various frequencies based on specified keywords were considered in the forecasting. The majority of the articles used monthly search query data in modeling and forecasting and only a few ones, including Pan et al. (2012) and Bangwayo-Skeete and Skeete (2015), used weekly search engine data directly without transformation for forecasting. Daily search data may be suitable for forecasting tourist attractions, such as the Forbidden City, China (Huang et al., 2017) and hotel occupancy (Pan et al., 2012). High-frequency search engine data can also be used directly for forecasting low-frequency tourism demand with a mixed data-sampling framework (Bangwayo-Skeete & Skeete, 2015; Song & Liu, 2017). Therefore, high-frequency search engine data can be used to forecast the tourism demand variable at a low frequency through the MIDAS approach, which improves the efficiency of the modeling and forecasting.

In terms of the keywords used for selecting search engine data, the reviewed articles relied primarily on domain knowledge related to tourism demand and the search query index under a specific category. For example, the first work by Pan et al. (2012) in this field used five tourist-related keywords from Google Trends to predict hotel room demand. The previous work by Choi and Varian (2012) adopted one subcategory on Hong Kong vacation destinations in Google Trends to conduct forecasting. The follow-up research improved and extended the number of keywords to reflect the comprehensive aspects of tourist activities. For example, Li et al. (2017) used 46 keywords to collect Baidu search data and predicted tourism demand in Beijing, China. Law et al. (2019) adopted 211 Google and 45 Baidu keywords to obtain search engine data and predicted tourism demand in Macau. Such a strategy is expected to cover available keywords concerning all aspects of tourists' interests on the destination (Law et al., 2019).

The methods for data selection, relevant data extraction, processing, and prediction are developed to incorporate search engine data further into the forecasting task (Brynjolfsson et al., 2016). Given that search engine data are massive and contain copious information, researchers should select the most useful information to unsure accurate forecasting through the use of selection and shrinkage methods (Song, Park, & Liu, 2019). The reviewed articles suggested that methods based on econometrics and deep learning have been developed to select and extract search engine data. Song and Liu (2017) proposed utilizing the least absolute shrinkage and selection operator and factor model to select search engine data. Li et al. (2017) and Li, Xu, et al. (2018) also adopted the principle component analysis (PCA) to reduce the dimensions of search engine data. Li et al. (2017) applied a generalized dynamic factor model to shrink the data and extract the most representative information of search engine data. For the latter, Law et al. (2019) applied the deep-learning technique to extract the most important data to improve forecasting. These two types of methods have different focuses. Methods, such as PCA or factor models tend to construct several indices that represent information in the search engine data, whereas methods for feature engineering suits the selection and extraction of search data from numerous potential variables.

Table 1
Summary of literature on tourism forecasting with Internet data.

Author (Year)	Internet data	Categorization	Forecasting methods	Forecasting context
Choi and Varian (2012)	Google Trends	Search engine	ARX	Visitor arrivals in Hong Kong from nine countries
Pan et al. (2012)	Google Trends	Search engine	Naïve, AR, ARX, ARMA, ARIMAX, **ARMAX**, ADL, TVP, VAR	Demand for hotel rooms in Charleston, SC
Orsi and Geneletti (2013)	Geotagged photographs	Social media	GM	Visitor flows in the Dolomites UNESCO World Heritage Site, Italy
Yang, Pan, and Song (2014)	Web traffic volume	Web traffic	**ARMAX**, ARMA, TAR	Hotel demand for a destination Charleston, South Carolina
Artola, Pinto, and Garcia (2015)	Google query index	Search engine	ARIMA, **ARIMAX**	Tourism inflows into Spain
Bangwayo-Skeete and Skeete (2015)	Google search data	Search engine	**AR-MIDAS**, AR, SARIMA	Tourism demand in five popular tourist destinations in the Caribbean
Kim, Lim, and Brymer (2015)	Social media reviews from multi-sources	Multi-source	MR	Performance of hotels in the U.S.
Yang, Pan, Evans, and Lv (2015)	Baidu, Google	Multi-source	ARMA, **ARMAX**	Tourist arrivals to Hainan, China
Yang, Tang, Luo, and Law (2015)	Web GIS	Web traffic	LR, PPR, ANN, SVR, BR	Performance of hotel locations in Beijing, China
Gunter and Önder (2016)	Web traffic data from Google Analytics	Web traffic	BVAR, FAVAR, VAR, BFVAR, **MA**, ETS, Naïve-1, **FCM**	Tourist arrivals to Vienna
Pan and Yang (2017)	Search queries Web traffic	Multi-source	**ARMAX**, MSDR	Hotel occupancy for a destination
Rivera (2016)	Google Trends	Search engine	**DLM**, SARIMA, HW, SNAIVE	Hotel non-resident registrations number in Puerto Rico
Huang, Zhang, and Ding (2017)	Baidu search data	Search engine	ARMA, **ADL**	Tourist arrivals to the Forbidden City, China
Li et al. (2017)	Baidu search data	Search engine	AR, **ARX**	Tourist arrivals in Beijing, China
Miah, Vu, Gammack, and McGrath (2017)	Photo-sharing social media sites, Flickr	Social media	LM, QM, EM	Future tourism demand for Melbourne
Önder (2017a)	Geotagged photos on Flickr	Social media	CA	Multi-destination trips in Austria
Önder (2017b)	Google trends web and image index	Multi-source	**ADL**, HW, Naïve-1	Tourist arrivals to Vienna, Barcelona, Austria, and Belgium
Park, Lee, and Song (2017)	Google trends	Search engine	ARIMA, SARIMA, SARIMAI, HW, **SARIMAX**	Japanese tourist inflow to South Korea
Peng, Liu, Wang, and Gu (2017)	Baidu search data	Search engine	HE-TDC	Tourism visitors in the Jiuzhai Valley scenic area, China
Zhang, Huang, Li, and Law (2017)	Baidu search data	Search engine	BA-SVR	Tourist arrivals to Hainan, China
Camacho and Pacce (2018)	Google search volume	Search engine	DFM	Tourist arrivals in Spain
Chang, Tsai, and Chiang (2018)	Social media from Twitter and Yelp	Social media	SITA	Hotel recommendation performance user preferences
Chong, Khong, Ma, McCabe, and Wang (2018)	Online reviews	Social media	SEM	Tourists' planning decisions
Dergiades, Mavragani, and Pan (2018)	Google trends	Search engine	VAR	Tourist arrivals in Cyprus
Li, Chen, et al. (2018)	Baidu data	Search engine	ARIMA, VAR, PCA-VAR, BPNN, PCA-BPNN, **PCA-ADE-BPNN**	Beijing in-bound tourist volume
Liu, Tseng, and Tseng (2018)	Baidu index combined with other variables	Multi-source	VAR	Tourism destination arrivals in Guizhou, China
Lv, Peng, and Wang (2018)	Baidu index and Google trends	Multi-source	SARIMA, MLR, SVR, SLFN, ESN, LSTM, **SAEN**	Tourism demand in America, Hainan, Beijing, and Jiuzhaigou (a scenic spot in China)
Ma, Xiang, Du, and Fan (2018)	Social media photos	Social media	DL	Helpfulness of hotel reviews
Tang (2018)	Baidu search engine	Search engine	ARIMA, SARIMA, HW, SARIMAI, **SARIMAX**	Chinese tourists to Thailand
Toral (2018)	Online reviews	Social media	K-nearest neighbor	Unique attributes of tourist destinations
Afzaal, Usman, and Fong (2019)	TripAdvisor and OpenTable	Multi-source	ASA	Sentiment classification of online tourist reviews
Bigne, Oltra, and Andreu (2019)	Twitter tweets and retweets	Social media	ANN	Hotel occupancy in short-break holidays for Spanish
Bokelmann and Lessmann (2019)	Google trends	Search engine	SARIMA, **DLM**	Short-term tourism demand in German holidays
Clark et al. (2019)	Google trends	Search engine	AR, **Google trends model**	Visitation in U.S. national parks
Colladon, Guardabascio, and Innarella (2019)	TripAdvisor travel forum, Google trends	Multi-source	AR, **FAAR**, **FABM**, BM (with exogenous variables)	International airport arrivals of seven major European capital cities
Emili, Figini, and Guizzardi (2019)	Google trends	Search engine	DM	Tourist arrivals and overnight stays in Florence, Milan, Catania, Germany, and the United Kingdom
Gai-Tzur, Bar-Lev, and Shiftna (2019)	Questions and Answer forums, TripAdvisor	Social media	Multidimensional model	Tourists' transport-related information
Hu and Song (2019)	Google trends	Search engine	**ANN**, ARIMA, ADL	Short-haul travel from Hong Kong to Macau
Huarng and Yu (2019)	Google trends	Search engine	HA	Taiwan tourism demand
Law, Li, Fong, and Han (2019)	Google trends	Search engine	DL, Naïve, SVR, ANN, ARIMA, **ARIMAX**	Macau tourist arrivals
Li, Lu, Liang, and Wang (2019)	Baidu index	Search engine	**FOA-BP**, GA-BP, PSO-BP	Tourist arrivals to a scenic area in China (Mount Huangshan)
Liu et al. (2019b)	Baidu index	Search engine	VAR	Tourist arrivals to Tianmu lake in China
		Multi-source	**KELM**, ARIMAX, ANN, LSSVR	Tourist arrivals in Beijing

(*continued on next page*)

Table 1 (*continued*)

Author (Year)	Internet data	Categorization	Forecasting methods	Forecasting context
Sun, Wei, Tsui, and Wang (2019)	Google trends, Baidu index			
Starosta, Budz, and Krutwig (2019)	Online media sentiments	Social media	Bivariate regression	Tourist arrivals in popular tourist destinations for Europeans
Volchek, Liu, Song, and Buhalis (2019)	Google trends	Search engine	**SARMAX-MIDAS**, ANN, Naïve, Seasonal Naïve, SARMA, SARMAX	Number of visits to five London museums
Wen, Liu, and Song (2019)	Google trends	Search engine	**Hybrid models,** ARIMA, ARIMAX, NAR, NARX	Tourist arrivals in Hong Kong from Mainland China
Zhang, Pu, and Wang (2019)	Baidu index	Search engine	**LSTM**, DBN, BPNN, C-LSTM	Hotel accommodation demand

Note. The method in bold indicates the best forecasting performance with Internet data.

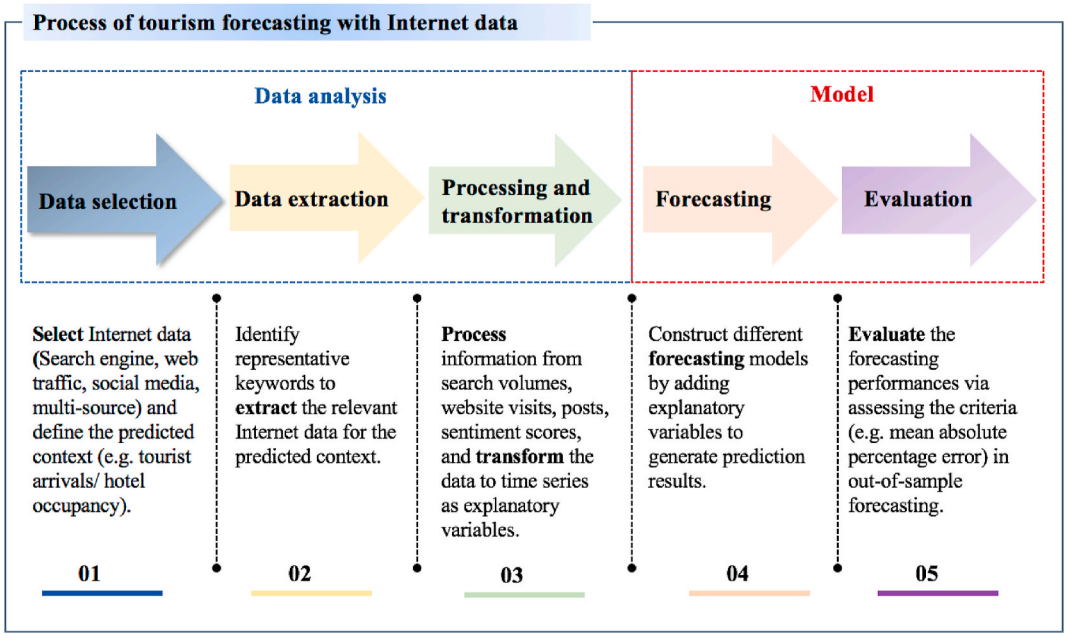

Fig. 3. Process of tourism forecasting with Internet data.

Time series, econometrics, artificial intelligence, deep learning, and hybrid models are introduced to model search engine data for forecasting. AR, ARMA, and ARIMA are among the most popularly used time series models, accounting for 44% of the reviewed articles. Such time series models are simple and effective and consider search engine data with a lag as explanatory variables. The lag orders of search engine data should be determined based on the Akaike information criterion (Li et al., 2017). Such models are also known as the benchmark that forecasts tourism demand with search engine data (Sun et al., 2019; Yang et al., 2015a). Increasingly advanced econometric models have been incorporated into forecasting tasks to improve accuracy further (Camacho & Pacce, 2018; Gunter & Önder, 2016; Orsi & Geneletti, 2013). These econometric models, including ADL, TVP, DFM, FAAR, Bayesian-FAVAR, and GM, are used in tourism forecasting and accounted for 28% of the reviewed studies. The above models address forecasting with search engine data under the same time-frequency. However, a MIDAS model can work when search engine data and the predicted variable, such as tourism arrivals, have different frequencies. Two reviewed studies under this category used such a model (Bangwayo-Skeete & Skeete, 2015; Volchek et al., 2019).

Artificial intelligence models used in the reviewed articles (16%) include artificial neural network and support vector regression, which are also known as "black boxes" in forecasting (Hu & Song, 2019; Li, Xu, et al., 2018; Li et al., 2019; Zhang et al., 2017). Li, Xu, et al. (2018) demonstrated that search engine data with PCA and a backpropagation neural network can outperform other forecasting models. Law et al. (2019) proposed that deep-learning-based models can outperform support vector regression and artificial neural network in the forecasting of Macau tourist arrivals with search engine data. The advantages of such deep-learning methods may rely on the built-in feature engineering ability when addressing large predictors (Law et al., 2019). Two reviewed articles (8%) used the long-short-term-memory models in forecasting under this category. Wen et al. (2019) combined the ARIMA and nonlinear ANN models to forecast tourist arrivals in Hong Kong from Mainland China with search engine data. The linear and nonlinear features of models can be utilized fully to enhance forecasting.

3.2. *Forecasting with web traffic data*

Web traffic data generally represent the raw number of visits to a website and indicate tourists' potential interests, which can be considered as explanatory variables to forecast tourist arrivals (Yang, Pan, & Song, 2014). Therefore, web traffic is also structured data. Yang, Pan, and Song (2014) pointed out that website visits are the step that follows the search for information through search engines and such data can have stronger predictive ability than search engine data. When tourists plan a vacation, they typically use search engines, such as Google and Baidu, to retrieve relevant information, and search results on destination management organizations (DMO) can be found on the first page (Gunter & Önder, 2016). Therefore, web traffic data from a DMO can predict tourism and hotel demands.

Two reviewed articles used web traffic data from a DMO to forecast tourism and hotel demands. Both collected data from Google Analytics, a free tool by Google. Yang, Pan, and Song (2014) applied two types of

web traffic volume data, namely, visitors and visits, that represented weekly data covering the 21st week of 2007 to the 17th week of 2011. The study examined whether web traffic data of a DMO can improve the forecasting accuracy of hotel rooms and occupancy rates in Charleston, US. The findings demonstrated that the ARMA model with web traffic data could reduce short-run forecasting errors significantly. As the first work using web traffic data for tourism/hotel forecasting, Pan et al. (2012) contributed to the literature by demonstrating that web traffic data can improve forecasting through predictive power.

Gunter and Önder (2016) used 10 web traffic data series from the Viennese DMO website to forecast tourist arrivals in Vienna. The reviewed study conducted a univariate benchmark and combined Bayesian VAR models and found that models performed differently between short and long horizons. Web traffic data used in this article are more comprehensive as compared to those of Pan et al. (2012). Along with *visitors* and *visits*, Gunter and Önder (2016) also employed indicators, such as *average session duration (in seconds)* and *returning visitors,* that reflected the time users spent on a specific session and the estimation of returning visitors.

Tourism forecasting research with web traffic data is still generally limited. However, such data can improve forecasting effectively. Time series and econometric models are used mainly in forecasting. Such data must be obtained from Google Analytics after seeking permission from DMO websites, which may entail further challenges in data collection.

3.3. Forecasting with social media data

Leung et al. (2013) suggested that social media has a significant effect on the tourism system by providing travelers with abundant channels to share through forums, blogs, microblogs, social networking, and photo and video sharing sites. These user-generated social media data are unstructured in contrast to a search engines and web traffic data (Chen et al., 2012). Social media-related research has been studied extensively, particularly on how online reviews affect tourist behaviors and hotel performance (Zhang et al., 2016; Ye et al., 2009). In terms of the research on tourism and hospitality forecasting, 10 of the reviewed studies adopted social media data for forecasting and 9 articles were published from 2017 to 2019. In general, the social media data used include texts and images, which appeared in the questions and answers in forums, user-provided reviews, and geotagged photos from various social media sites, such as Twitter, TripAdvisor, Priceline, Hotels.com, Expedia, Flickr, and Yelp (Ma et al., 2018). Specifically, text- and image-based forecasting research accounted for 60% and 40%, respectively. The former included posts, reposts, and sentiments reflected in reviews, whereas the latter was dominated by user photos.

Text-mining-based approaches, including clustering, topic modeling, and visualization, are utilized to address unstructured textual data and extract relevant keywords, generate important features, and identify patterns (Toral, Martínez-Torres, & Gonzalez-Rodriguez, 2018). Toral et al. (2018) identified unique attributes of tourist destinations through a bag of words from reviews collected from online communities and adopted machine-learning methods to train classifiers. Their findings demonstrated that the unique attributes were the best predictors of tourist destinations. Similarly, Bigné, Oltra, and Andreu (2019) used a text-mining approach to extract important information from tweets on Twitter to examine how DMOs' Twitter activities affect hotel occupancy forecasting. Based on the contents and using text mining, they classified tweets into different categories, including those related to events, tourist attractions, socialization, and commercials. The findings identified five influential factors that can predict hotel occupancy rate for a destination, which include the number of user retweets, user replies, event-related tweets, tourist attraction-related tweets, and retweets by DMOs. In contrast to search engine and web traffic data, researchers should extract useful structured data from social media data with techniques, such as text mining.

Along with the extraction of textual information in social media data, the sentiment reflected in the texts was also considered in the existing literature. Starosta et al. (2019) computed the sentiment indices for popular destinations for Europeans, which reflected positive or negative attitudes of online media toward the destinations. With the machine-learning method, the sentiment index showed online sentiment as a time series that can be incorporated to predict tourist arrivals. Their findings revealed that the sentiment index constructed from social media data can serve as an explanatory variable to predict tourism demand. Chang et al. (2018) used social media data from Twitter and combined it with data from Yelp to improve hotel recommendation performance. They discovered that earlier posts on Twitter and Yelp can measure tourists' posting behavior. Afzaa, Usman, and Fong (2018) proposed an aspect-based sentiment classification to identify tourist reviews from social media platforms and that the extracted aspects can improve prediction accuracy.

For image-based social media data, machine-learning techniques, such as clustering, are employed to analyze geographical photos. For example, Miah et al. (2017) adopted a density-based cluster algorithm to identify tourists' photos on Flickr and forecast future tourism demand in Melbourne. Onder (2017) used cluster analysis to classify trips in Austria based on geotagged photos on Flickr. Orsi and Geneletti (2013) adopted geotagged photos and a gravity model to estimate tourist flows and identify popular locations. When tourists provided photos accompanied by review texts in social media sites, Ma et al. (2018) argued that the combination of photos and texts in Yelp and TripAdvisor can predict review helpfulness more accurately than text and image only. Deep-learning models were used to predict review helpfulness. Thus, the photos provided by the users can help other users decide on the value of information on social media sites.

Unstructured social media data, including texts and images, are transformed into structured time series. We can use forecasting techniques, such as econometric and artificial intelligence models, to predict tourism demand. In particular, 60% of the reviewed articles used econometric forecasting models, and 40% adopted artificial intelligence approaches. The artificial neural network approach was popular in connection to forecasting tourism demand.

3.4. Forecasting with multi-source data

Pan and Yang (2017) argued that additional big data sets should be integrated to improve forecasting accuracy further. In the nine reviewed articles that combined multiple data sources for forecasting, three categories of combination can be observed, namely, two different search engine data (3/9), one search engine data and other variables including web traffic data and social media data (3/9), and different social media data (3/9).

Given that the search engine data were from different sources that represent tourist behaviors in various markets, the combination of search engines can represent tourists from different markets. Sun et al. (2018) suggested that integrated data from the Baidu and Google index can improve forecasting accuracy significantly. Yang et al. (2015a) argued that Baidu data can forecast tourism demand in China more accurately than Google data because of the large market share of the Baidu search engine. Concurrently, the Google search data can be used effectively in forecasting tourism demand in countries that mainly use the Google search engine. Pan and Yang (2017) combined Google search and web search traffic data to forecast the hotel occupancy rate for a destination and demonstrated that the combination of different data sources can reduce forecast errors and improve forecasting accuracy. Search engine data can also be combined with other variables, including structured statistical data, such as the number of daily tickets and semi-structured data. Semi-structured data indicate data that do not follow a conventional database system, which may be in the form of structured data that are not organized in relational databases such as tables (Hashem et al., 2015). Liu et al. (2018) suggested that weather, temperature, and calendar information, which include weekends and

public holidays are semi-structured data. Accordingly, Liu et al. (2018) combined Baidu search engine data and the above variables to forecast tourist arrivals and discovered that not all variables can contribute to the significant improvement of forecasting accuracy.

Internet data generated from multiple sources represent the diverse aspects of tourists' behavior and such data can entail the different ways of improving forecasting accuracy. Social media data generated from different platforms, including Twitter, Facebook, TripAdvisor, and Yelp, can have different effects. Xiang, Du, Ma, and Fan (2017) compared social media data quality in different platforms, such as TripAdvisor, Expedia, and Yelp, and uncovered huge differences in how the hotel industry is represented. Social media platforms vary in quality and focus on different user segments. TripAdvisor and Yelp appeared to have more helpful reviews than Expedia (Xiang et al., 2017). Colladon et al. (2019) examined whether social media data from the TripAdvisor travel forum combined with Google Trends data can forecast tourism demand accurately. Researchers collected several semantic variables based on posts written in the English language and constructed variables that reflected the social network in online communities, including social structure, interactivity, and rotating leadership. Sentiment and complexity indices were built to measure the used language in the posts. Their findings indicated that combined data contributed to forecasting differently and that the language complexity and social network-related variable can improve the forecasting of international airport arrivals.

From the perspective of forecasting models, time series and econometric models continued to occupy dominant roles in the forecasting with combined data sources accounting for 67% in the reviewed articles. The commonly used models include ARMA, ARMAX, ADL, VAR, and FAAR (Fronzetti et al., 2019; Pan & Yang, 2017; Yang et al., 2015a). Machine learning, such as kernel extreme learning machine and stacked autoencoder, were used to model and forecast tourism demand (Lv et al., 2018; Sun et al., 2019). Sun et al. (2019) argued that the proposed machine learning model was most effective and stable in the forecasting experiments and robustness analysis.

3.5. Summary of Internet data sources

Table 2 summarizes the review findings of four Internet data sources: search engine, web traffic, social media, and multi-source data, including their advantages, limitations, results, managerial implications, and unresolved issues that would point out potential directions in tourism forecasting with Internet data.

Generally, Internet data can be viewed as new driving factors that represent tourists' attention, interests, and sentiments that influence tourism forecasting (Li et al., 2017; Song et al., 2019). The adoption of Internet data in tourism forecasting will have managerial implications that can support DMOs in gaining a better understanding of tourist behaviors, allocate resources, and form timely decisions to improve tourism demand (Ma et al., 2018; Song et al., 2019). However, limitations on the data quality because of search keywords selection, noise, or irrelevant information in social media data, theoretical interpretations, and robust empirical examinations remain (Song et al., 2019). In particular, existing studies on social media and multi-source data are still limited, and researchers need to explore what data sources can be integrated from a rigorous and robust perspective. Better methodologies to process huge unstructured Internet data should be proposed. The issues on how many Internet data sources could provide the most precise forecasting accuracy compared to a single data source should also be addressed.

4. Discussion and implications

The developments of Internet technology and big data have had a considerable effect on the tourism and hospitality industry (Leung et al., 2013; Li, Xu, et al., 2018; Wu, Song, & Shen, 2017). In tourism forecasting, Song, Liu, and Park (2019) indicated the use of Internet data has been a driving factor in the development of forecasting models. Our review considers four categories of Internet data used in tourism forecasting and synthesizes the methodological developments in recent decades. Fig. 4 summarizes the number of reviewed papers based on data sources and forecasting models over the past eight years. Several general trends on the development of tourism forecasting with Internet data are observed in our review.

4.1. General trends in forecasting with Internet data

First, tourism-forecasting research with search engine data became increasingly popular during the studied periods. The first article on tourism forecasting with Internet big data collected search engine data from Google (Choi & Varian, 2012; Pan et al., 2012) and sought to

Table 2
Summary of Internet data sources.

	Search engine	Web traffic	Social media	Multi-source data
Advantages	Low cost of data collection Structured time-series data Favorable applicability to different research topics	Retrieved from Google Analytics Structured time-series data Timely data at weekly and monthly frequencies	Generated from different social media platforms Contain useful information on tourist behaviors	Abundant data sources that reflect tourist behaviors comprehensively More robust than a single data source
Limitations	Noises and irrelevant data may affect the accuracy Require better data shrinkage methods and avoid estimation issues	Data are not open to the public Several indicators are unavailable Difficult to apply the data to more forecasting contexts	Require tools, such as crawlers to obtain data Needs to improve data quality through text mining, and sentiment analysis methods	Time costs in collecting multi-source data Contains more noises than single-source data Requires higher efficient methods to analyze the data
Results	Reflects tourists' attention or interests and improves forecasting accuracy	Reflects DMO websites' visits/popularity and improves forecasting accuracy	Reflects tourists' positive or negative sentiments and improves forecasting accuracy	Comprehensively reflects tourists' attention, interests, and sentiments and improves forecasting accuracy
Managerial implications	To support management sectors to understand tourists' attention promptly To provide a new driving factor to improve forecasting accuracy	To support DMO websites to allocate the resources, reduce cost, detect tourist behavior and make decisions	To provide managers or marketers with timely responses and make decisions based on the sentiments in reviews and photos	To better understand tourist behaviors from different platforms Make precise strategies to improve forecasting
Unresolved issues/ potential directions	To ensure accuracy and coverage of selected search keywords To propose better methods in data extraction and model estimation	To explore other available web traffic data To examine forecasting performances of web traffic data in more contexts	To remove noise and extract useful index from social media data To illustrate the theoretical support for social media data that can affect forecasting	To explore what data sources can be integrated To verify empirical applications of multi-source data in the forecasting

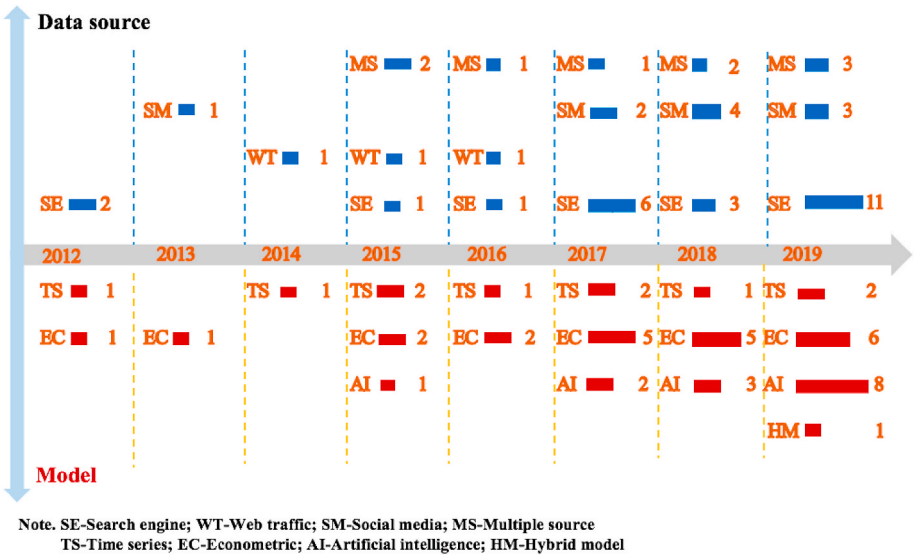

Fig. 4. Evaluation of Internet data sources and forecasting models (2012–2019).

incorporate this type of Internet data into other research contexts in tourism. Relevant articles published in the past three years in this category accounted for 83%. Existing studies have demonstrated the incorporation of search engine data can improve forecasting accuracy (Dinis, Breda, Costa, & Pacheco, 2019; Liu, Liu, Wang, & Pan, 2019a). Search engine data are obtained from Google or Baidu, and the structured time-series data can be applied to forecast different outcomes in the fields of tourism and hospitality (Yang et al., 2015a). However, the selection of search keywords determines the data quality that can affect forecasting accuracy (Geva et al., 2017). Researchers need to achieve a balance between the coverage of keywords and forecasting performance. Therefore, complicated and advanced methods on search engine data are constantly evolving. For example, machine learning and deep learning-based models were adopted to extract useful features, reduce the complexity of the forecasting models, and improve performance (Law et al., 2019; Sun et al., 2019).

Second, our analysis indicated that tourism-forecasting research with social media and multi-source data has also been gaining increasing attention. Social media research in tourism and hospitality has been recognized widely over the past decades (Leung et al., 2013; Li & Law, 2020). Advanced methods, including text mining and sentiment analysis, have been used to transform the unstructured social media data into the structured data series. The selection of social media platforms, extraction, and representation of social media data are important steps that need to be taken before forecasting. Researchers seek to explore how the most representative social media data can be used to improve the forecasting accuracy of tourism demands. Advancements in machine learning methods can contribute to understanding the text and image information on social media platforms (Li & Jiao, 2020). In comparison with structured search engine data, most social media data are generally unstructured and complex. Therefore, relatively abundant social media data, including structured data (e.g. the number of reviews) and unstructured texts and images, can be provided for timely tourism forecasting.

Third, from the perspective of forecasting models, econometric models are still dominant and artificial intelligence methods have also been gaining popularity. Time series models, such as ARMA, are considered as the benchmark model in the forecasting. Researchers have also used deep-learning methods to improve forecasting accuracy, particularly when massive Internet data with nonlinear information are employed (Wen et al., 2019). Law et al. (2019) demonstrated that a deep-leaning framework reduced forecasting errors in comparison with support vector regression and artificial neural networks. However, given the differences among forecasting contexts, identifying which models have the highest forecasting accuracy is difficult. Song and Li (2008) argued that no models can outperform others. Similarly, our analysis did not compare forecasting accuracy among these different models. Time series, econometrics, and artificial intelligence models have their advantages. Methodological developments indicate that hybrid models will likely be incorporated to combine these advantages in future research (Jiao & Chen, 2019).

4.2. Implications for tourism management

Search engine, web traffic, social media, and multi-source data contribute to tourism forecasting by providing new explanatory variables that reflect the different behaviors of tourists in a more timely and effective manner. The incorporation of such new data sources and the relevant data modeling techniques improve forecasting accuracy and provide practical important implications for destinations to form better decisions on the tourism market (Li & Jiao, 2020; Song et al., 2019; Wu et al., 2017).

Search engine data are the most popular among Internet data sources used in tourism forecasting by researchers and practitioners. Search engine data are easy to access through Google or Baidu and can be used and verified directly by researchers, practitioners, and policymakers in analyzing tourism demand (Li et al., 2017). Such data have implications for tourism management, which provide timely data sources to understand tourists' attention and improve the precision of tourism forecasting. Our observations indicate social media data have been gaining popularity in tourism forecasting. Social media data that can be collected from various social media platforms contain more abundant information about tourist behaviors as compared with search engine data (Onder, 2017). The incorporation of social media data can represent tourists' specific positive or negative attitudes and sentiments that can be used to understand their responses and formulate correct decisions for tourism management sectors.

Our findings suggest multi-source data have also been attracting increasing attention from researchers and industries. Researchers have been collecting multi-source data for tourism forecasting to enrich their data sources further and overcome the limitations of a single one (Liu et al., 2018; Sun et al., 2019). For example, data sources that include search engine and social media data can not only reflect tourists' attention but also their sentiments and attitudes (Colladon et al., 2019). Multi-source data from search engines, web traffic, and different social media will contribute to methodological developments and advanced

analytical techniques can be used to modeling the data. Therefore, the applications of multi-source data could potentially provide significant implications for destinations where policy-makers can formulate timely strategies based on different perspectives of the tourism market.

4.3. Future directions

Given the differences among the four kinds of Internet data, several concerns must be addressed in the modeling and forecasting of tourism and hospitality, which require profound investigation in future research. First, the quality of search engine data including the coverage and accuracy of keywords determine the performances of tourism forecasting (Geva et al., 2017). Hence, researchers should arguably focus on proposing more robust and rigorous search keywords selection methods that can largely reduce the noise and irrelevant information of search engine data and analyze massive Internet data effectively. Recent developments in machine learning will help provide better solutions for these questions. For example, Law et al. (2019) adopted a deep-learning framework to conduct feature selection from massive search engine data.

Second, although social media data have certain advantages containing rich information reflecting tourists' positive or negative sentiments, issues on the quality of social media data, such as online reviews are also critical for tourism forecasting research (Xiang et al., 2017). As indicated from review findings, the quality of social media data, including the authenticity of data and sentiments, varies across different platforms. Therefore, methodological developments in natural language processing and sentiment analysis should be considered thoroughly in future research. Moreover, social media data, such as texts and images appearing in reviews, are unstructured and need to be transformed into a structured time series in forecasting. Generally, the transformation procedure is one distinct difference from search engine data. Researchers have devoted efforts to obtaining structured time series from texts and images by using text-mining and sentiment analysis (Starosta et al., 2019). Advancements in natural language processing can be used in tourism forecasting with social media and multi-source data.

Although Internet data, such as search engines and social media, have been widely discussed in previous studies, the theoretical foundation of each kind of data used in the forecasting remains unclear. As suggested by Song et al. (2019), more efforts should be devoted to exploring the economic implications behind Internet data sources and the theoretical support on how these data can influence tourism demand. Given that the popularity of various data generated in Internet platforms continues to grow, tourism industries should attach more importance to the adoption of such data in accurate forecasting. Considering real-time Internet data as supplements for traditional data sources used in the tourism forecasting is important. Future tourism forecasting could become timelier and transform from monthly to daily and hourly forecasts (Li & Jiao, 2020). Various Internet data can also be adopted to enhance the performance and reduce bias because each type of data has its disadvantages. Geva et al. (2017) indicated that search engine data cannot reflect users' sentiments compared with social media data. Certain search engine data, such as Google data, are typically at an aggregate level and raw searches are not available to researchers. Accordingly, the combinations of various Internet data, including Google, Baidu, and different social media data can make the data sources more comprehensive by capturing more specific information of each data source (Luo, Zhang, & Duan, 2013). More accurate and timely forecasting results would be achieved by incorporating different Internet data and appropriate models.

5. Conclusions

The increasing popularity of Internet big data has resulted in its development as a kind of new data source for tourism demand forecasting. Song et al. (2019) pointed out that big data have a huge potential for improving forecasting accuracy. Researchers have developed different competitive forecasting models, including time series, econometrics, artificial intelligence, and hybrid models to achieve higher accuracy. Internet data generated from various platforms have different formats in structured or unstructured forms that require different modeling techniques in tourism forecasting. In reviewing the 47 relevant tourism forecasting articles that incorporated one or several Internet data sources from 2012 to 2019, this study summarized the general trends of tourism forecasting using four types of Internet data: search engine, web traffic, social media, and multi-source data. We find tourism forecasting research with search engine data has been gaining popularity during the studied periods, while forecasting research with social media and multi-source data have been attracting considerable attention. Time series and econometric models are more commonly used in tourism forecasting as compared to artificial intelligence methods. Hence, advancements in natural language processing could be used in tourism forecasting with social media and multi-source data.

We synthesize the findings and present how different forecasting models, including time series, econometrics, artificial intelligence, and hybrid methods are adopted in addressing the different Internet data. Discussions of this study are useful for researchers interested in tourism forecasting with Internet data. The superiority of the combination of different Internet data in tourism forecasting is expected because each type of Internet data has certain advantages and disadvantages, which may lead to estimation bias and forecasting errors. The combination of search engines, web traffic, and social media data will considerably enrich the data source and enhance performance. More hybrid models with multi-source data are also expected for the modeling and forecasting of tourism demand. However, the incorporation of more semi-structured and unstructured Internet data would entail modeling and forecasting challenges, which will require contributions from diverse multidisciplinary fields, such as computer science and mathematics (Song et al., 2019).

This study contributes to the literature by providing an overview of the findings of a recent popular research topic on tourism forecasting with various Internet data. Although big data in tourism have been discussed widely, the application of Internet data in tourism forecasting is still in its infancy. Hence, we contribute to the academia and industry by presenting research trends in extant studies and implications for future tourism forecasting research with Internet data. This study has several limitations that may serve as future directions. We only analyzed full-length peer-reviewed articles in the English language from academic journals. Reviews, conference articles, editorial, abstracts, and letters were excluded, which could have limited the sample size. Hence, future research should consider more sources, such as leading conferences in the fields of tourism and hospitality to provide more insights into this topic.

Author contribution statements

Xin Li developed the research idea and conducted the systematic review. Xin Li launched the literature selection and analyzed the literature with Gang Xie's advice. Xin Li wrote the manuscript. Gang Xie, Rob Law, and Shouyang Wang reviewed, revised, and proofread the whole manuscript.

Impact statement

The increasing growth of Internet applications has generated a huge amount of data, and the Internet data are considered important supplements to traditional data for government, industries and academia. Although big data in tourism have been discussed widely, the application of Internet data in tourism forecasting is still in its infancy. This study comprehensively reviewed the research state of different Internet data used in tourism forecasting, and traced the methodological development of Internet data: search engine, web traffic, social media, and

multiple sources. This study addressed the superiority of combined Internet data sources and hybrid forecasting models in incorporating semi-structured and unstructured data in practices. Findings can provide useful implications to accurately apply Internet data to achieve a better forecasting performance. Therefore, the study may shed light on the appropriate usage of data and modelling techniques for precise tourism forecasting for government, destination management organizations and tourism industries.

Declarations of competing interest

None.

Acknowledgments

The authors acknowledge the support of research fund from the National Natural Science Foundation of China (No. 71601021) and the Fundamental Research Funds for the Central Universities (No. FRF-TP-19-067A1).

References

Afzaal, M., Usman, M., & Fong, A. (2019). Predictive aspect-based sentiment classification of online tourist reviews. *Journal of Information Science, 45*(3), 341–363.
Artola, C., Pinto, F., & de Pedraza García, P. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower, 36*(1), 103–116.
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.
Berthon, P. R., Pitt, L. F., Plangger, K., & Shapiro, D. (2012). Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy. *Business Horizons, 55*(3), 261–271.
Bigné, E., Oltra, E., & Andreu, L. (2019). Harnessing stakeholder input on twitter: A case study of short breaks in Spanish tourist cities. *Tourism Management, 71*, 490–503.
Bokelmann, B., & Lessmann, S. (2019). Spurious patterns in Google Trends data - an analysis of the effects on tourism demand forecasting in Germany. *Tourism Management, 75*, 1–12.
Brynjolfsson, E., Geva, T., & Reichman, S. (2016). Crowd-Squared: Amplifying the predictive power of search trend data. *MIS Quarterly, 40*(4), 941–961.
Camacho, M. (2018). Forecasting travellers in Spain with Google's searches. *Tourism Economics, 4*(4), 434–448.
Camacho, M., & Pacce, M. J. (2018). Forecasting travellers in Spain with Google' s searches. *Tourism Economics, 24*(4), 434–448.
Chang, J. H., Tsai, C. E., & Chiang, J. H. (2018). Using heterogeneous social media as auxiliary information to improve hotel recommendation performance. *IEEE Access, 6*, 42647–42660.
Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *The Economic Record, 88*, 2–9.
Chong, A. Y. L., Khong, K. W., Ma, T., McCabe, S., & Wang, Y. (2018). Analyzing key influences of tourists' acceptance of online reviews in travel decisions. *Internet Research, 28*(3), 564–586.
Clark, M., Wilkins, E. J., Dagan, D. T., Powell, R., Sharp, R. L., & Hillis, V. (2019). Bringing forecasting into the future: Using Google to predict visitation in U.S. national parks. *Journal of Environmental Management, 243*, 88–94.
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.
Dergiades, T., Mavragani, E., & Pan, B. (2018). Google Trends and tourists' arrivals: Emerging biases and proposed corrections. *Tourism Management, 66*, 108–120.
Dinis, G., Breda, Z., Costa, C., & Pacheco, O. (2019). Google trends in tourism and hospitality research: A systematic literature review. *Journal of Hospitality and Tourism Technology, 10*(4), 747–763.
Emili, S., Figini, P., & Guizzardi, A. (2019). Modelling international monthly tourism demand at the micro destination level with climate indicators and web-traffic data. *Tourism Economics, 26*(7), 1129–1151.
Gal-Tzur, A., Bar-Lev, S., & Shiftan, Y. (2019). Using question & answer forums as a platform for improving transport-related information for tourists. *Journal of Travel Research, 59*(7), 1221–1237.
Gunter, U., & Önder, I. (2016). Forecasting city arrivals with Google analytics. *Annals of Tourism Research, 61*, 199–212.
Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115.
Hu, M., & Song, H. (2019). Data source combination for tourism demand forecasting. *Tourism Economics, 26*(7), 1248–1265.
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.
Huarng, K. H., & Yu, T. H. K. (2019). Application of Google trends to forecast tourism demand. *Journal of Internet Technology, 20*(4), 1273–1280.
Jiao, E. X., & Chen, J. L. (2019). Tourism forecasting: A review of methodological developments over the last decade. *Tourism Economics, 25*(3), 469–492.
Kim, W. G., Lim, H., & Brymer, R. A. (2015). The effectiveness of managing social media on hotel performance. *International Journal of Hospitality Management, 44*, 165–171.
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.
Leung, D., Law, R., Van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing, 30*(1–2), 3–22.
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, G., & Jiao, X. Y. (2020). Tourism forecasting research: A perspective article. *Tourism Review, 75*(1), 263–266.
Li, X., & Law, R. (2020). Network analysis of big data research in tourism. *Tourism Management Perspectives, 33*, 100608. https://doi.org/10.1016/j.tmp.2019.100608
Li, K., Lu, W., Liang, C., & Wang, B. (2019). Intelligence in tourism management: A hybrid FOA-BP method on daily tourism demand forecasting with web search data. *Mathematics, 7*, 531.
Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57–66.
Liu, H., Liu, Y., Wang, Y. L., & Pan, C. C. (2019a). Hot topics and emerging trends in tourism forecasting research: A scientometric review. *Tourism Economics, 25*(3), 448–468.
Liu, Y. Y., Tseng, F. M., & Tseng, Y. H. (2018). Big data analytics for forecasting tourism destination arrivals with the applied Vector Autoregression model. *Technological Forecasting and Social Change, 130*, 123–134.
Liu, P., Zhang, H., Zhang, J., Sun, Y., & Qiu, M. (2019b). Spatial-temporal response patterns of tourist flow under impulse pre-trip information search: From online to arrival. *Tourism Management, 73*, 105–114.
Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management, 68*, 301–323.
Luo, X., Zhang, J., & Duan, W. (2013). Social media and firm equity value. *Information Systems Research, 24*(1), 146–163.
Lv, S. X., Peng, L., & Wang, L. (2018). Stacked autoencoder with echo-state regression for tourism demand forecasting using search query data. *Applied Soft Computing Journal, 73*, 119–133.
Ma, Y., Xiang, Z., Du, Q., & Fan, W. (2018). Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. *International Journal of Hospitality Management, 71*, 120–131.
Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information Management, 54*(6), 771–785.
Önder, I. (2017a). Classifying multi-destination trips in Austria with big data. *Tourism Management Perspectives, 21*, 54–58.
Önder, I. (2017b). Forecasting tourism demand with Google trends: Accuracy comparison of countries versus cities. *International Journal of Tourism Research, 19*(6), 648–660.
Orsi, F., & Geneletti, D. (2013). Using geotagged photographs and GIS analysis to estimate visitor flows in natural areas. *Journal for Nature Conservation, 21*(5), 359–368.
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.
Pan, B., & Yang, Y. (2017). Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research, 56*(7), 957–970.
Park, S., Lee, J., & Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. *Journal of Travel & Tourism Marketing, 34*(3), 357–368.
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.
Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management, 57*, 12–20.
Song, H., & Liu, H. (2017). Predicting tourist demand using big data. In *Analytics in smart tourism design* (pp. 13–29). Cham, Switzerland: Springer.
Song, H., Qiu, R. T., & Park, J. (2019). A review of research on tourism demand forecasting. *Annals of Tourism Research, 75*, 338–362.
Starosta, K., Budz, S., & Krutwig, M. (2019). The impact of German-speaking online media on tourist arrivals in popular tourist destinations for Europeans. *Applied Economics, 51*(14), 1558–1573.
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.
Tang, J. (2018). Evaluation of the forecast models of Chinese tourists to Thailand based on search engine attention: A case study of Baidu. *Wireless Personal Communications, 102*(4), 3825–3833.
Toral, S. L., Martínez-Torres, M. R., & Gonzalez-Rodriguez, M. R. (2018). Identification of the unique attributes of tourist destinations from online reviews. *Journal of Travel Research, 57*(7), 908–919.
Volchek, K., Liu, A., Song, H., & Buhalis, D. (2019). Forecasting tourist arrivals at attractions: Search engine empowered methodologies. *Tourism Economics, 25*(3), 425–447.
Wen, L., Liu, C., & Song, H. (2019). Forecasting tourism demand using search query data: A hybrid modelling approach. *Tourism Economics, 25*(3), 309–329.

Wu, D. C. G., Song, H., & Shen, S. J. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management, 29*(1), 507–529.
Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51–65.
Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015a). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397.
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.
Yang, Y., Tang, J., Luo, H., & Law, R. (2015b). Hotel location evaluation: A combination of machine learning tools and web GIS. *International Journal of Hospitality Management, 47*, 14–24.
Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*(3), 6527–6535.
Zhang, B., Huang, X., Li, N., & Law, R. (2017). A novel hybrid model for tourist volume forecasting incorporating search engine data. *Asia Pacific Journal of Tourism Research, 22*(3), 245–254.
Zhang, B., Pu, Y., Wang, Y., & Li, J. (2019). Forecasting hotel accommodation demand based on LSTM model incorporating Internet search index. *Sustainability, 11*(17), 4708.
Zhang, Z., Zhang, Z., & Yang, Y. (2016). The power of expert identity: How website-recognized expert reviews influence travelers' online rating behavior. *Tourism Management, 55*, 15–24.

Xin Li, Ph.D., is an Associate Professor at School of Economics and Management, University of Science and Technology Beijing. Her research interests are big data analytics, econometric modeling, data mining and forecasting.

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.

Gang Xie, Ph.D., is an Associate professor at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China. His research interests include econometrics, artificial intelligence and forecasting.

Shouyang Wang, Ph.D., is a Professor at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China. His research interests include decision analysis, risk management, economic analysis and forecasting.