Hengyun Li, Qian Wang, Lingyan Zhang & Danting Cai (2022) Big Data in China Tourism Research: A Systematic Review of Publications from English Journals, Journal of China Tourism Research, 18:3, 453-471

BIG DATA RESEARCH IN CHINA TOURISM: A SYSTEMATIC LITERAURE REVIEW

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

Qian Wang*
Shenzhen Research Institute,
The Hong Kong Polytechnic University,
Shenzhen, China
Email: wangq779@nenu.edu.cn
* Corresponding Author

Lingyan Zhang
Shenzhen Research Institute,
The Hong Kong Polytechnic University,
Shenzhen, China
Email: lingyan.zhang@u.nus.edu

Danting Cai, Ph.D. Student
School of Hotel and Tourism Management,
The Hong Kong Polytechnic University,
Hong Kong SAR, China
Email: dan-ting.cai@connect.polyu.hk

Declaration of competing interest
None

Acknowledgments
The authors acknowledge the support of research fund from the National Natural Science Foundation of China (71902169).

Big data in China tourism research: A systematic review of publications from international journals

Abstract: Expansive data, collected and stored at unprecedentedly granular levels, are transforming the tourism landscape. Big data has received considerable attention in tourism-related research; a growing number of scholars are exploring the latest applications to exploit this resource's underlying potential. Regular systematic reviews of the field's progress and academic trends are needed, yet few efforts have addressed the Chinese context. Accordingly, this paper assesses 66 tourism articles featuring big data analytics in past decades and summarizes the literature by data types and research topics. Current trends and future directions are discussed based on the findings.

Keywords: Big data; Tourism research; China; Literature review

摘要: 数字化时代中消费者产生的海量数据正在改变旅游业的面貌。相应地，大数据在旅游研究中受到了相当大的关注；越来越多的学者正在探索最新的应用，以开发这一资源的潜在潜力。尽管有学者对这一领域的进展和学术趋势进行了初步的系统审查，但没有单独针对中国情境下的研究的回顾。因此，本文评估了在中国情境下展开的、以大数据分析为特色的 66 篇旅游文章，并按数据类型和研究主题对文献进行了总结。基于我们的发现，本文讨论了当前的趋势和未来的方向。

关键词: 大数据；旅游研究；中国；文献综述

1. Introduction

Constant technological milestones have opened a new chapter of social life featuring increasing exposure to the digital environment, wherein countless structured and unstructured large-scale data are produced (Li et al., 2018). This environment also generates many types of big data. The most commonly used definition of big data is data that is large (volume), complex (diversity), and variable (velocity) (Lazer et al., 2021). Big data is gaining popularity in the social sciences (Pan & Yang, 2017), especially in the tourism sector, thanks to its potential to reveal underlying values and drive industry operational excellence. Consumer decision making is multidimensional; many choices rely on an array of information sources and objective conditions (Li, Pan, et al., 2017). This intricacy amplifies the shortcomings of traditional data collection methods (i.e., questionnaires and case studies), such that conclusions are drawn amid numerous constraints and control variables. Big data can inform solutions to associated problems given inherent advantages that have revolutionized the research ecology of tourism. Notable limitations of traditional methods (Lv et al., 2021) include overly small sample sizes, subjective bias, and sampling bias (Li, Pan, et al., 2017; Yang et al., 2015), while the sheer scale of big data allows for an unprecedented increase of robustness of the results that could not be achieved by traditional methods in the past (Law et al., 2019; Li, Hu, et al., 2020).

In the digital age, consumer trajectories and activities are being captured and stored in the form of user data by various service providers, which grant access to relevant stakeholders to accurately yet intuitively translate the data into consumer preferences (Lazer & Radford, 2017). A change of popular research topics has been observed due to the availability of such big data. Scholars are empowered to dive into the minds of consumers and make predictions (e.g., demand forecasting and route planning) that is of great benefits for the planning and management of cities and scenic spots from a managerial perspective (Law et al., 2019; Li, Hu, et al., 2020). Considering how big data has altered the trajectory of tourism studies, this study is keen to understand the issue more fully, particularly in China, in view of its importance as a tourist generating source globally (UNWTO, 2011, 2020). A report released by the World Tourism Cities Federation documented 12.31 billion global tourism arrivals in 2019, of which China accounted for 5.88 billion or 47.7% (WTCF, 2021). The global tourism industry continues to face severe challenges from the COVID-19 pandemic. Meanwhile, Chinese tourism has witnessed a robust post-pandemic recovery surpassing all other markets (Statista, 2021). China's most distinctive developmental feature is the intertwining of its state and market; this environment is unlike any other region in the world and frames the Chinese tourism market as a unique setting worthy of focused research (Wen & Wu, 2020; Xu et al., 2008). Through a case study of a Chinese tourist attraction, Li, Hu, et al. (2020) demonstrated how big data has improved the performance of tourism demand forecasting and offered valuable practical implications. Besides, Xu (2017) reported that big data is playing an increasingly important role in Chinese tourism. Thus, an overview of tourism big data research in this context is especially warranted. Such an effort will summarize the evolution of big data applications in China for comparison with big data research in the international context.

Scholars have attempted to review the existing research on big data. Li et al. (2018) pioneered this attempt by classifying big data into user-generated content data, device data and transaction data, and sorting out the themes of existing research according to the different data types. Subsequently, a bibliometric review by Li and Law (2020) summarized research on the topic of big data from 2008-2017, providing a comprehensive overview of research trends, popular areas and thematic clusters through visual network analysis. In addition to adopting a holistic perspective, some scholars have focused on one specific topical area vertically in their review. For example, Li, Law, et al. (2021) focused specifically on the Internet data and its application in tourism demand forecasting studies, while Rahmadian et al. (2021) reviewed the application of big data in sustainable tourism. Despite the remarkable contributions of many scholars, one direction that still merits additional discussion is that no studies have yet focused on a review of Big Data research in the context of a particular country or region. China—the world's largest tourist source market—has not received sufficient academic attention, and country-specific research on big data is lacking. Thus, a systematic review of the literature focusing on the Chinese context is necessary. Notably, Jin (2020) has collected papers published in Chinese journals and reviewed the research on Big Data at the Chinese level. Therefore, this study, in particular, focuses on Big Data research in the Chinese context by examining publications in international journals, with an aim to offer a more comprehensive perspective from the international research community (Bao et al., 2018).

The aim of this study is to provide a comprehensive review of tourism-related Big Data applications in China, by distinguishing relevant research by data type and discussing progress along relevant research lines. Relevant challenges and future research directions involving big data in the Chinese tourism industry are also outlined. To the best of our knowledge, this paper is the first review of tourism Big Data research in this national context.

2. Methodology

The systematic search strategy entailed several steps. Similar to other systematic review articles (Li, Law, et al., 2021), two widely adopted databases – Web Of Science and Google Scholar - were employed to search the literature. In terms of journal selection, this study did not apply the filtrations by targeting specific journals with an aim to ensure the integrity of the review, but applied the exclusion method through the content of the publications. Then, keyword screening were performed to identify potentially relevant publications. Keywords included but were not limited to (big data OR search engine data OR user-generated data OR social media data) and (China OR Chinese) and (tourism OR hospitality) to ensure maximum literature coverage. Next, eligibility criteria were established for further screening. Only peer-reviewed full-length papers and English-language publications were kept in this step; research notes, reviews, conference papers, and books were excluded. Therefore, all literature on the topic of tourism and big data upon search results was retained before proceeding to abstract and full text checking. Finally, two research assistants were invited to manually identify articles of little content relevance by reading the abstracts and full text of all included sources. In addition, each article's reference list was checked to ensure that no signature sources were inadvertently overlooked in our review process. Sixty-six papers were obtained as of September 2021. **Figure 1** summarized the progress of literature selection. Of the literature sources collected, 45 were published after 2018, accounting for 67% of the sample.

The following analysis falls into two parts: (1) To provide a statistical overview of the literature visually, several descriptive analyses were performed in R using Bibliometrix package. Specifically, annual publications, relevant sources, high frequency keywords, temporal distribution and thematic changes of the literature were presented; (2) A systematic literature review was adopted as the methodological approach by classifying previous studies according to data types and research topics with an aim to answer the formulated questions - What are the main data types that were leveraged in the big data research in China tourism? What are the research topics that were investigated based on each data type?

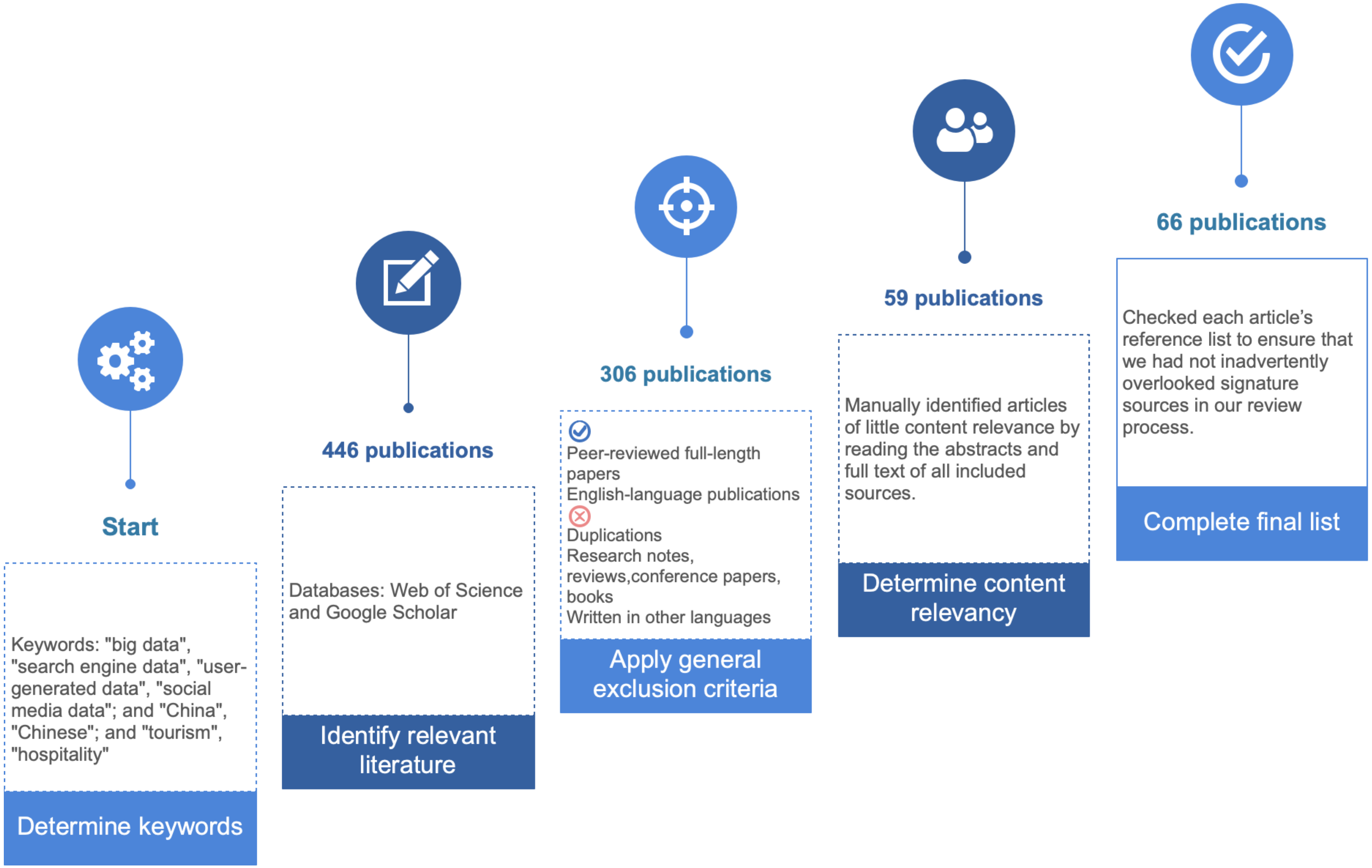

Figure 1. Literature Selection Progress

3. Results

3.1 Results of descriptive analysis

Figure 2 illustrates the distribution of annual academic output on big data in tourism research. The first paper in the Chinese context was published in 2010. The left-skewed distribution reflects growing scholarly interest over time, with a peak observed in 2020. As noted, the sources obtained for 2021 consisted of a partial year: only those published before September were collected. The majority of articles (22) appeared in *Tourism Management*, followed by *Journal of Travel Research* (9) and *Asia Pacific Journal of Tourism Research* (7); see **Table 1 and Figure 3**. Words that include “China”, “tourist behavior”, “tourist destination”, “forecasting method” and “tourism market” appear most frequently (**Figure 4**), suggesting the popular research topics of interest in the sample. In addition, the keyword network was grouped into three clusters by keyword co-occurrence analysis (**Figure 5**). Cluster 1 includes publications that are lexically linked to tourists’ experiences and satisfactions, such as tourists’ satisfaction towards hospitality by examining the electronic word-of-mouth. Cluster 2 contains publications related to tourism demand forecasting, featuring the usage of time-series data and artificial intelligence model to predict the tourists’ arrival volumes. Cluster 3 consists of publications related to tourists’ mobility patterns and destination image. These two topics are closely related and fall into one cluster since majority of relevant research share the similarity of emphasizing on tourists’ onsite behavior or perceptions in a specific destination context. **Figure 5** also conveys the temporal evolution of keywords co-occurrence based on the lexical network, illustrating the change of mainstream research topics over years. The study horizon was chronologically broken into sub-periods represented by colors from bluish to orange. Keywords related to tourism demand forecasting are mainly colored in greenish and orange, implying the growing popularity of such topics from 2019 onwards.

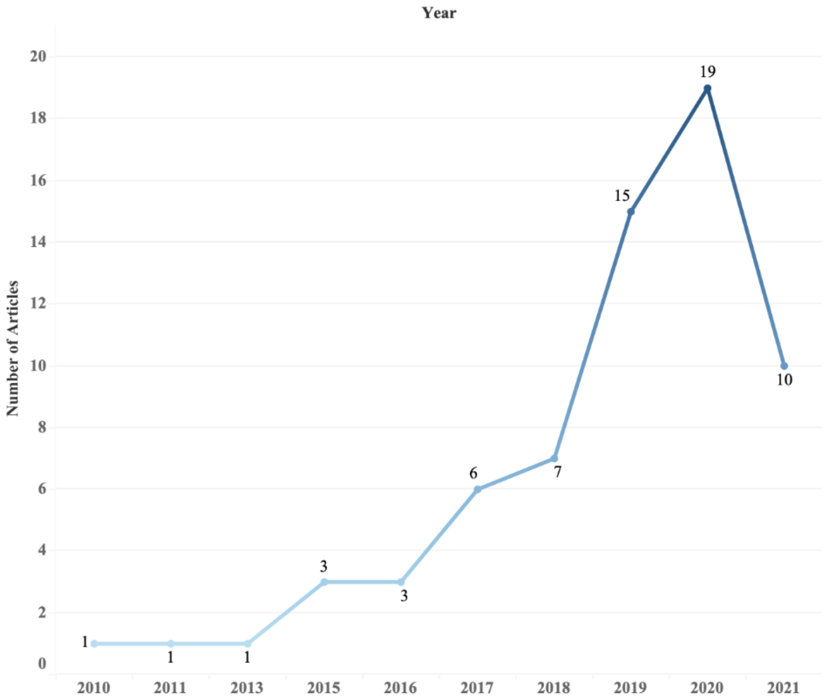

Figure 2. Progression of Big Data Research in China Tourism Research

	Year										
Source Title	**2010**	**2011**	**2013**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**Total**
TOURISM MANAGEMENT			1	2	2	3	3	7	3	1	22
JOURNAL OF TRAVEL RESEARCH							3	1	3	2	9
ASIA PACIFIC JOURNAL OF TOURISM RESEARCH						2		1	4		7
INTERNATIONAL JOURNAL OF CONTEMPORARY HOSPITALITY MANAGEMENT									1	3	4
TOURISM ECONOMICS								2	1		3
INTERNATIONAL JOURNAL OF TOURISM RESEARCH		1							1	1	3
ANNALS OF TOURISM RESEARCH								1	2		3
TOURISM MANAGEMENT PERSPECTIVES								1		1	2
JOURNAL OF TRAVEL & TOURISM MARKETING							1	1			2
INTERNATIONAL JOURNAL OF HOSPITALITY MANAGEMENT				1					1		2
CURRENT ISSUES IN TOURISM									1	1	2
JOURNAL OF VACATION MARKETING					1						1
JOURNAL OF HOSPITALITY MARKETING & MANAGEMENT									1		1
JOURNAL OF DESTINATION MARKETING & MANAGEMENT										1	1
ISPRS INTERNATIONAL JOURNAL OF GEO-INFORMATION								1			1
INTERNATIONAL JOURNAL OF HOSPITALITY & TOURISM ADMINISTRATION	1										1
INFORMATION TECHNOLOGY & TOURISM						1					1
IEEE ACCESS									1		1
Total	1	1	1	3	3	6	7	15	19	10	66

Table 1. Most Relevant Sources by Year

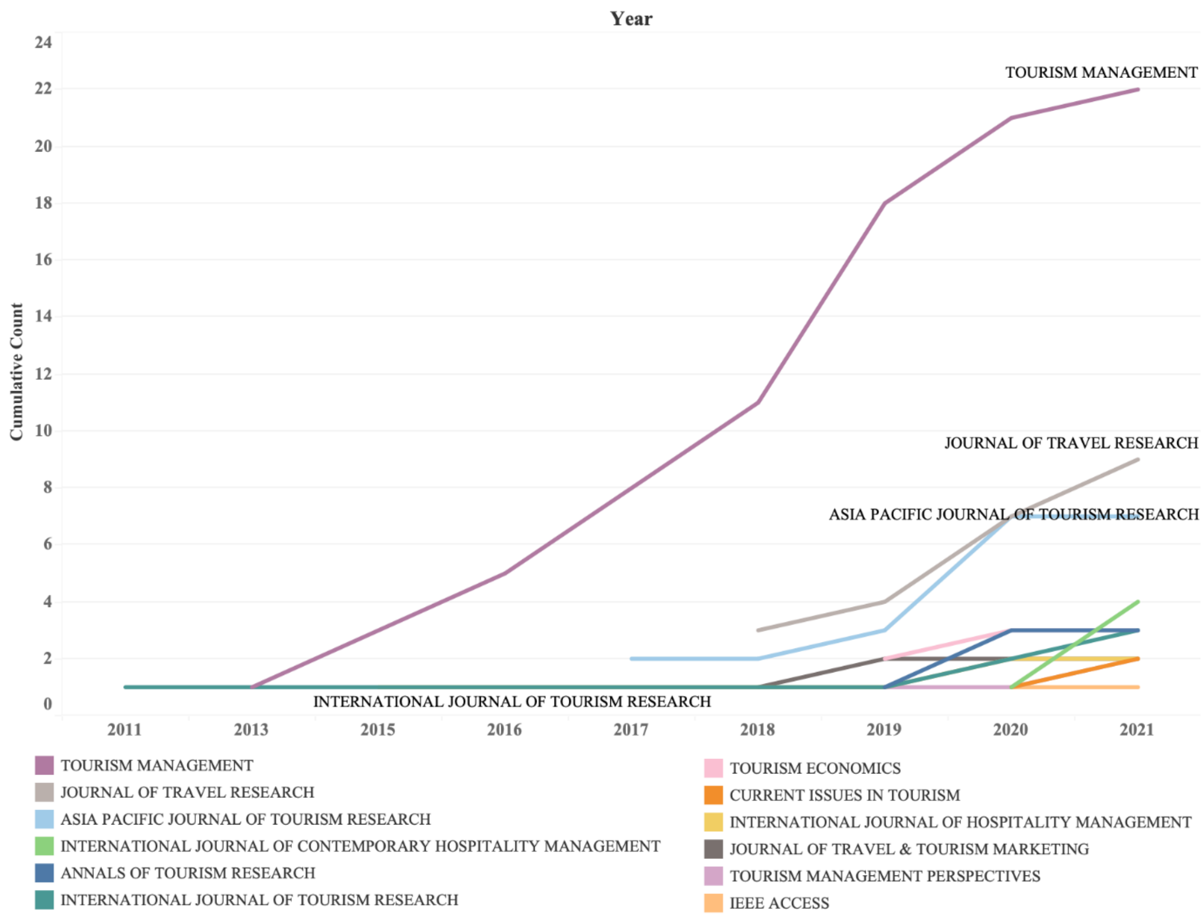

Figure 4. Keywords Frequency

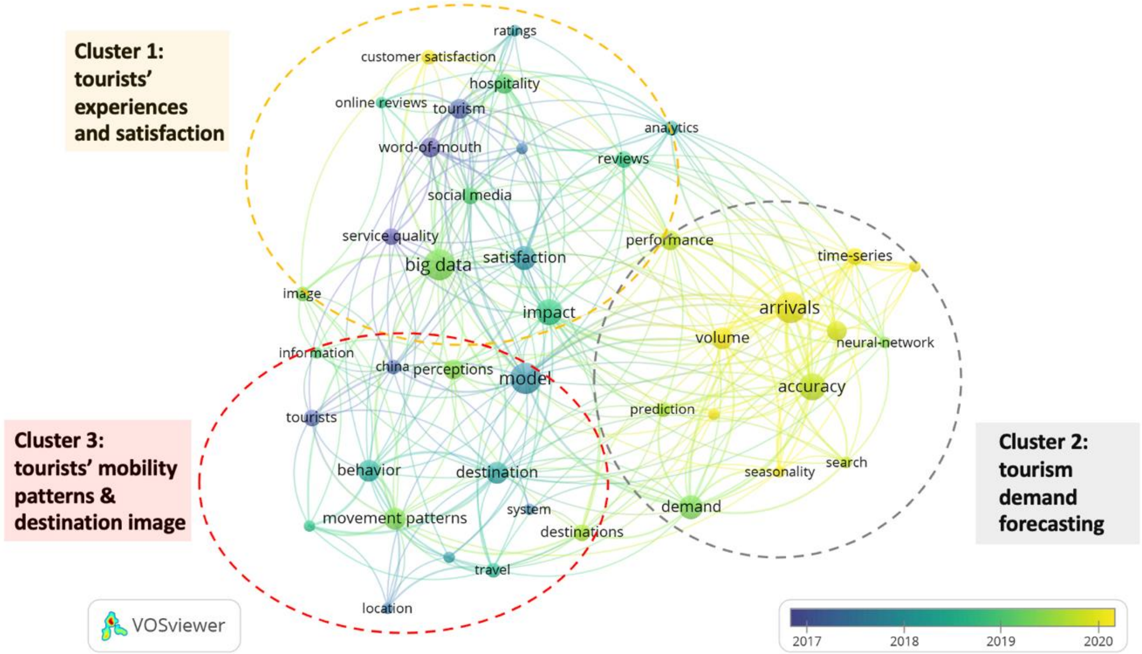

Figure 5. Lexical and Temporal Network

Several intriguing patterns emerged from analysing the locations profiled across sources. First, authors most often focused on popular tourist cities with typical Chinese characteristics, mainly in coastal areas (e.g., Hong Kong, Macau, Beijing, Shanghai, and Sanya). Hong Kong and Beijing were the most popular sites, with 17 and 16 studies unfolding in either city respectively. Studies conducted in less well-known and geographically remote locations, such as inland cities, were relatively rare (**Figure 6**). Second, mainland cities garnered particular attention owing to their ongoing economic development, strong cultural output, and growing regional influence: authors' focus from 2015 onwards clearly shifted from Hong Kong, Macau, and Taiwan to mainland cities (**Figure 7**). Third, among studies conducted at the individual tourist attraction level, most concerned globally popular scenic spots such as the Forbidden City and Summer Palace.

Figure 6. Map of Focal Sites in Studies

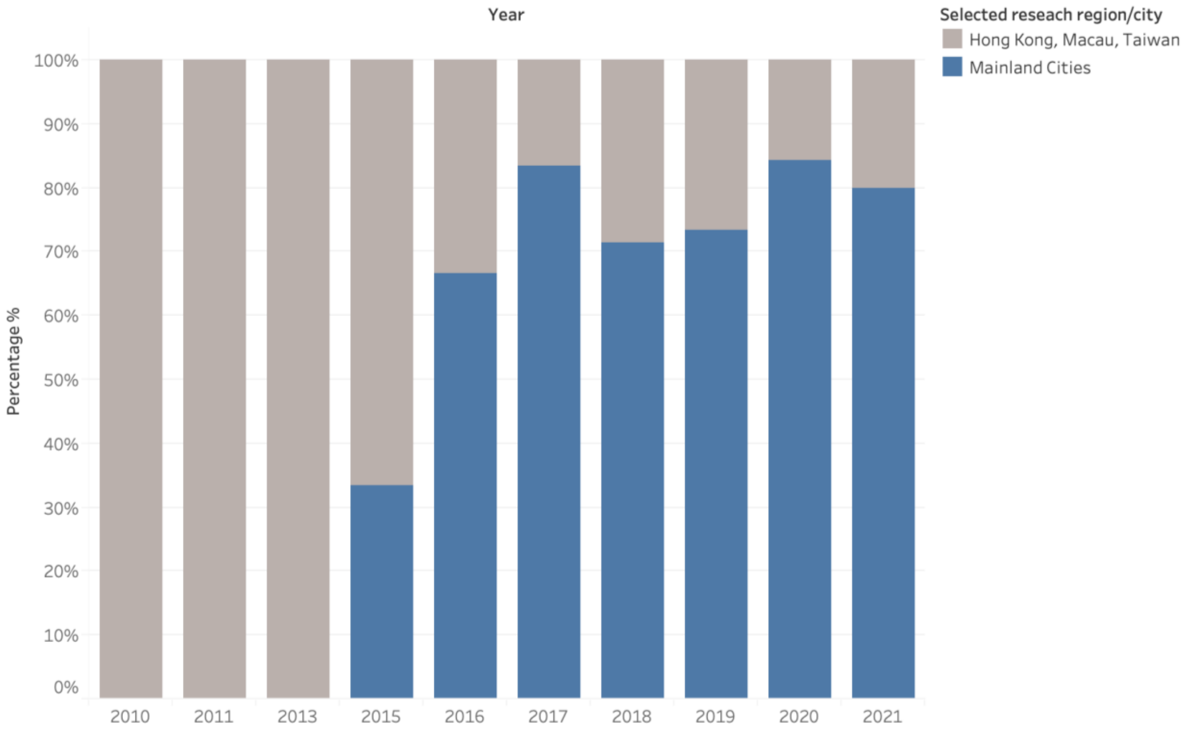

Figure 7. Percentage of Mainland Cities Studied per Year

3.2 Results of the systematic review

Tourists generate data throughout the travel process. Technological advances have enabled a tremendous volume of data to be captured, stored, and analyzed at unprecedentedly granular levels. Many types of data have been integrated in tourism as evidenced by our sample. The maturity of computer-based data processing techniques has allowed researchers to leverage applicable data. In this section, findings based on the content of the literature classified by data types and topics were presented.

The type of data under consideration informs how and in which contexts researchers can exploit it. Different data types were employed across the sample to solve tourism-related questions; therefore, studies can be compiled into specific categories: (1) search engine data (i.e., consumer queries on Baidu and Google search engines, usually consisting of phrases or keywords); (2) user-generated data (i.e., text data and photo data, both actively shared by tourists on various social media platforms); (3) GPS data (i.e., tourists' location data); and (4) mixed data (i.e., any two or more types of big data mentioned above).

In general, major topics within the chosen literature on tourism in China covered four research streams: (1) tourism demand forecasting, (2) tourists' experiences and satisfaction, (3) tourists' mobility patterns, and (4) destination image. Summarization of the literature on the bases of data types and research topics was presented in Table 2. User-generated data and search engine data were most popular; the former was more often used to analyze visitors' experiences and satisfaction, while search engine data were frequently applied in tourism demand forecasting. In the following sections, discussions were provided based on each data type.

Categorized Data Type	Research Focus	Year 2010	2011	2013	2015	2016	2017	2018	2019	2020	2021	Total
User-generated data	Tourist experience & satisfaction			1	1	2	2	1	4	5	1	17
	Destination image	1	1					1	3	2	1	9
	Tourists' mobility patterns				1		1	2	1	2		7
	Tourism demand forecasting							1	1		1	3
Search engine data	Tourism demand forecasting				1	1	2	1	4	5	2	16
GPS data	Tourists' mobility patterns						1	1	2		1	5
Mixed data	Tourism demand forecasting									1	2	3
	Tourists' mobility patterns									1		1
	Tourist experience & satisfaction									1		1
	Destination image									1		1
Other types of data	Tourists' mobility patterns										2	2
	Tourist experience & satisfaction									1		1
Total		1	1	1	3	3	6	7	15	19	10	66

Table 2. Research Foci over Time

3.2.1 Search engine data

In the age of information, visitors are increasingly inclined to conduct research via search engines. The traces of these interactions can be stored as a new kind of online data—search engine data (Yang et al., 2015). These data can reveal trends in users' interests, behavior, and attitudes; they also contain information helpful for predicting tourists' behavior (Li, Law, et al., 2021). In addition, search data empowers scholars to conduct real-time monitoring and generate on-the-spot insights, transcending the boundaries of delayed information dissemination associated with traditional statistics (Huang et al., 2017). Therefore, search indices such as the Baidu search index and Google search index are now widely used as predictors in tourism demand forecasting.

In the Chinese context, tourism research based on search engine data has mainly addressed demand forecasting. A large number of scholars have documented the exemplary performance of search indices in improving models' forecasting accuracy (Hu & Song, 2019; Zhang et al., 2017). Search engine data at different frequencies have been used to predict demand. Gawlik et al. (2011) were the first to incorporate a monthly search index into a hotel forecasting model to improve forecasting performance. Scholars have since continued to introduce search data at higher frequencies, such as weekly (Li, Hu, et al., 2020) and daily (Hu et al., 2021; Zhang, Li, et al., 2020). Other researchers have compared the performance of search indices on multiple platforms, including Baidu and Google (Sun et al., 2019; Yang et al., 2015).

The keyword selection of search engine data is another important topic. Choosing highly relevant keywords with high coverage can greatly influence tourism demand forecasting performance (Li, Li, et al., 2021). A common approach involves extracting initial keywords for tourism demand forecasting based on prior domain knowledge and then extending the pool by incorporating related keywords. Finally, to reduce noise, many approaches have been proposed to extract valid information from a large amount of search data. First, many researchers have used the correlation values between tourism demand and keyword queries to identify the most typically used keywords for forecasting. For example, Yang et al. (2015) selected 20 queries as seed keywords based on the travel planning process. Next, they retrieved related queries and calculated the correlations between visits and different queries to finalize the keyword queries by retaining only those with the highest correlation values. Bi et al. (2020) determined initial search keywords based on domain knowledge (e.g., "name of the tourist attraction", "attraction weather", "attraction hotel", and "attraction airport") and then extended them by using Baidu's keyword auto-recommendation technology. Similar to Yang et al. (2015), Bi et al. (2020) only kept keyword queries with the highest correlation values regarding tourism demand. Second, principal component analysis (PCA) (Li, Pan, et al., 2017; Tang et al., 2021) is another practical approach to reduce both data dimensionality and noise in search engine data. To better represent the dynamics among all search queries and to minimize information loss, Li, Pan, et al. (2017) created a composite search index using a generalized dynamic factor model to acquire the most relevant information from search keyword queries. Third, some machine learning methods have been adopted to identify valid information (Li, Li, et al., 2021); examples include filter-based feature selection, recursive feature selection, genetic algorithm feature selection, and random forest feature selection.

3.2.2 User-generated data

The prevalence of social media has fueled the explosion of user-generated content, with massive amounts of content being produced. Today's service providers tend to gain competitive advantages by targeting a vertical and then achieving supremacy. Therefore, content shared on various platforms is diverse and distinct. Three main categories of online platforms featuring user-generated content exist, which served as data sources for most studies in the sample. The first category is social network platforms, referring to broadly used sites such as Facebook and Twitter where people share updates and engage with friends and communities. The second category includes professional review sites, such as TripAdvisor and Dianpin.com, where individuals share experiences and opinions about products or services they have consumed; their input can potentially influence prospective customers' decision making. The third category is composed of content-sharing sites, such as YouTube, TikTok, and Flickr, which host a variety of user-generated content created for different purposes.

User-generated content can be divided into many categories depending on the purpose of the consumer sharing it and the platform. Generally, however, when distinguishing user-generated content in terms of format, such data are either text-based or image-based.

3.2.2.1 Text-based data

Text-based data constitute the dominant data source in tourism research, as early online platforms usually stemmed from microblogs featuring text-centric posts. Travel blogs and online reviews are two prime data sources in this category. Travel blogs were one of the first forms of textual data to be used in tourism studies; nearly all relevant research has referred to this type of data to evaluate destination image. As one of the earliest attempts, Leung et al. (2011) summarized tourists' destination image perceptions of Hong Kong by analyzing travel blogs posted on Ctrip.com. Law and Cheung (2010) gathered 120 travel blogs from different websites to analyze mainland tourists' image perceptions of Hong Kong as well. In addition to evaluating tourists' image perceptions of specific destinations, scholars have used larger data samples to determine the destination images of multiple cities. Wang et al. (2019) collected 140,000 travel blogs covering 20 cities and found that tourists held distinct image perceptions of different locations. For instance, tourists visiting cities such as Beijing and Shanghai valued culture, entertainment, and advanced facilities, whereas those visiting Guilin and Chengdu valued natural scenery and the environment. Hao et al. (2021) collected more than 50,000 travel microblogs to assess mainland tourists' impressions of Hong Kong during a turbulent period. In terms of methodology, content analysis and sentiment analysis are the main approaches that researchers have employed to parse travel blog data.

Online reviews have represented another major text-based data source in recent years, supported by the rise of social media and professional review sites along with people's desire to be seen and heard. Word-of-mouth serves as a powerful tool to appeal to and influence fellow consumers. Although scholars have leveraged online reviews for multiple purposes, most have concentrated on destination image and the tourist experience. Jiang et al. (2021) analyzed more than 70,000 reviews from three platforms in mainland China (i.e., Baidu Travel, Tuniu, and Poor Travel) to uncover mainland tourists' image perceptions towards Hong Kong. Liu et al. (2019) derived Chinese tourists' image perceptions of Australia based on tourist-generated textual content and discerned discrepancies between Chinese visitors and international visitors.

The second research topic using online reviews relates to the tourist experience. Online reviews briefly document a consumer's journey and objectively reflect their overall travel experience. Thus, these reviews are valid cues of visitors' experiences in numerous settings, including restaurants (Jia, 2020), hotels (Luo et al., 2021; Ying et al., 2020), and destinations (Li, Tung, et al., 2017). Most studies have sought to discover factors that shape the tourist experience, including attributes related to climate (Liu et al., 2021), the environment (Fan et al., 2021; Zhang, Yang, et al., 2020), racial discrimination (Li, Li, et al., 2020), tourists' cultural backgrounds (Jia, 2020; Schuckert et al., 2015; Ying et al., 2020), hotels (Liu et al., 2017; Luo et al., 2021), and destination attributes (Li, Tung, et al., 2017). For example, Ying et al. (2020) explored Chinese tourists' preferred characteristics and desired services for accommodation based on review data from Ctrip.com and TripAdvisor for hotels in six major tourist cities. Li, Tung, et al. (2017) investigated consumers' satisfaction with four types of attractions by analyzing over 200,000 traveler reviews. Jia (2020) assumed a cross-cultural perspective when comparing the preferences of Chinese and American tourists towards a selected restaurant by referring to 2,448 user-generated reviews; results showed that Chinese tourists favored a quiet environment and high-quality food whereas American tourists were less concerned about whether the dining environment was crowded.

3.2.2.2 Image-based data

Photographs are an inherent aspect of tourism. Visual content is beginning to dominate textual content as an information source for several reasons. From a consumer perspective, posting photos is less time-consuming while accurately conveying one's desired message; from a research perspective, visitor-uploaded photographs contain far more information than text (Zhang et al., 2019). Images also show more than simply visual information—they contain temporal data, geographic data, and metadata. Each type provides additional information, thus acting as a rich resource for tourism studies while offering scholars a fresh perspective. Meanwhile, processing massive amounts of data enhances the robustness of researchers' results. Most studies in our sample that included image-based content involved tens of thousands of observations per dataset. The study findings were therefore more trustworthy from a statistical perspective than previous work that relied heavily on manual content analysis for hundreds of records.

More importantly, the development of technological tools, exemplified by artificial intelligence, has made it possible to analyze unstructured data such as image information; this task was previously unachievable. User-generated photos have been adopted in the Chinese tourism context to accomplish three aims: (1) to extract tourists' perceived destination image from user-generated images; (2) to reveal tourists' travel patterns and places of interest based on the geographical information contained in images; and (3) to forecast tourism demand using tourist-generated images.

The first type of research focuses on revealing tourists' perceived destination image. As mentioned above, photos can more richly reflect visitors' perceptions than text in this regard. Additionally, photographs not only contain clear locations but can also convey tourists' emotions, both of which can help scholars discern destination image. Zhang et al. (2019) analyzed international tourists' image perceptions of Beijing by collecting photos on Flickr and discovered that attractions bearing typical traditional Chinese cultural characteristics (e.g., the Forbidden City and the Great Wall) contributed to inbound tourists' image perceptions of Beijing; food was less important. Visitors from different source regions also possessed varying images of Beijing. Deng et al. (2019) similarly analyzed a dataset from Flickr and found that visitors from distinct cultural backgrounds held significantly different image perceptions of Shanghai.

The second research area reveals tourism patterns based on geographic information contained in pictures. Related studies have explored tourist segments' spatial-temporal behavior, including tourist routes and points of interest. For example, Vu et al. (2015) developed a new method based on geographic photos to evaluate tourists' spatial-temporal behavior. They took Hong Kong as an example to portray visitors' travel trajectories by analyzing inbound tourists' photos on Flickr, which indicated travelers' preferences for different attractions. Deng and Liu (2021) analyzed 15,000 photos posted on Instagram to determine points of interest among different tourists in Beijing. Zhang et al. (2019) used deep learning to scrutinize variations in the movement patterns of tourists from different source markets when visiting Beijing. Scholars have also analyzed disparities in tourists' travel patterns and preferences in relation to tourist attractions, such as museums (Vu et al., 2018) and temples (Leung et al., 2017).

The third research stream entails tourism demand forecasting using tourist-generated image data. This topic remains in its infancy. Demand forecasting intends to predict the number of future destination visitors based on photos generated by past visitors. Even though only one study in the sample used images to conduct forecasting, the core idea mirrored that of earlier forecasting studies: to make predictions by extracting metadata from photos. Chen et al. (2019) collected geotagged photos posted by tourists on Flickr and used these images to predict inbound passenger flows in Beijing. In particular, the authors used the P-DBSCAN clustering algorithm to analyze photos' metadata (e.g., user ID, latitude and longitude, location, time of day, and photo's title and label) to predict the number of visitors to Beijing. They further compared the prediction results with statistical yearbook data to demonstrate the accuracy of using photos to predict demand.

3.2.3 GPS data

GPS offers an effective technology with which to acquire tourist movement data thanks to advantages such as high accuracy and a high response rate (Shoval & Isaacson, 2007). GPS data can currently be collected from tourists in two ways: either by issuing tourists GPS devices to record their travel trajectories at scenic spots or by obtaining these data through cellphones. In the latter case, researchers can refer to signal data from cellphone operators and from Google Maps or Baidu Maps (Li et al., 2018).

GPS data have been applied in two key subject areas. The first is tourists' movement patterns. Essentially, individuals' travel patterns demonstrate changes in tourists' spatial-temporal behavior. Wu and Chen (2021) used GPS data to compare differences in the travel patterns of tourists and local residents at the Confucius Temple in Nanjing. GPS data from tourists visiting a zoo in Beijing also revealed significant differences in tourists' and local residents' travel patterns (Xu et al., 2020). Zhong et al. (2019) examined the travel patterns of tourists in Tibet and assessed the appeal of multiple attractions. In addition to individual tourists' travel patterns, group tourists' movement patterns have been considered as well: using Xi'an as an example, Zhao et al. (2018) demonstrated the significant effect of group size on travel patterns. The authors then analyzed group tourists' shared travel patterns along with identifying differences in tourists' travel patterns and preferences based on demographics.

Furthermore, GPS data can improve the prediction efficiency and accuracy of destination/attraction recommendations. Zheng et al. (2017) proposed a heuristic method to predict visitors' next location at an attraction by considering tourists' trajectories in Xiamen. They further used GPS data to evaluate this method's performance at the attraction. Relatedly, Zheng et al. (2018) proposed a method to predict similarities in tourists' trajectories by analyzing GPS data. By evaluating the differences between historical and current visitor trajectories, this approach was found to enhance visitors' experiences and destination recommendations. The authors verified its effectiveness at the Summer Palace.

4. Conclusion

Tourism is developing rapidly in the digital age as tourists' traces are recorded in novel ways. The resultant data present fresh directions and perspectives to investigate complicated tourism-related problems. This paper has surveyed the application of big data in tourism within the Chinese context via a systematic review of data types and research topics. The possible future of big data in tourism is summarized below.

4.1 Findings

A statistical analysis of the literature revealed that the application of big data in China's tourism industry remains immature. More studies are needed to fully exploit the potential of this type of data. Even so, several valuable conclusions can be drawn. The sample of 66 papers in this study featured three main types of big data (i.e., search engine data, user-generated data, and GPS data). Each type has unique merits and is suitable for certain topics. For example, search engine data are mainly used for tourism demand forecasting, whereas GPS data can depict tourists' movement patterns. Second, the chosen tourism studies leveraged big data to address four major topics: destination image, tourists' experiences and satisfaction, mobility patterns, and tourism demand forecasting. Third, scholars' interests are clearly shifting over time. Destination image was the first mainstream researching topic for which big data was adopted in the Chinese context, while movement patterns and demand forecasting have received more attention as of late. For instance, 23 papers involved demand forecasting in recent years. Studies' focal locations have also moved towards mainland cities, accompanied by a decline in research targeting areas such as Hong Kong, Macau, and Taiwan. Finally, cutting-edge analytical methods have led computer-based artificial intelligence methods to take over this subject area; traditional methods contingent on manual data processing are receding.

4.2 Challenges and future research directions

Although academic interest in big data tourism research in the Chinese context is expanding, the quality of big data remains a point of contention (Li et al., 2018). Reducing noise and adopting optimized analytical methods are equally important when addressing big data–related problems. Data quality directly affects the validity and robustness of findings. Data pre-processing is therefore crucial; denoising, complementing, and clustering can help to ensure high-quality data for analysis (Li, Ge, et al., 2020). Existing literature has devoted relatively little attention to this matter.

Future efforts in this area should seek to expand the domain's boundaries and enhance the value of big data in tourism research. Publications in Chinese language were excluded from the sample in this review. Subsequent studies can also analyze relevant articles from Chinese domestic journals to develop a better understanding of the current research progression. Still, this study points to several promising avenues for big data tourism studies in the Chinese context. First, scholars should incorporate video-based data into their research designs. It is found that the types of big data adopted thus far have been limited, especially in terms of user-generated content, which focuses on text- and image-based analysis. The popularity of short videos and advances in 5G networks have caused a growing number of videos to be posted on social media. These videos offer an additional resource to consider; they carry abundant information, and future work on tourism in China would benefit from investigating this data type.

Second, academics could fuse data from multiple sources to enhance the value of big data applications. Most studies so far have referred to a single data source to analyze specific travel behaviors. Travel is a naturally complex process through which tourists generate diverse data reflecting numerous behaviors and attitudes. Multi-source data can provide a multidimensional, multi-level perspective for tourism research, leading to a clearer understanding of tourists' activity patterns and behavioral characteristics that cannot be achieved with single-source data. For example, search engine data reflect tourists' pre-trip planning and preferences, while user-generated data reflect individuals' actual travel and attitudes. Yet the quality and structure of data vary by source. As such, handling massive volumes of multi-source heterogeneous data presents a series of obstacles: fusion, migration, and complementation of multi-source information; and the representation, modeling, and collaboration of heterogeneous information. Addressing these challenges is paramount to enlarging the added value of big data applications.

Third, most research on tourism in China has adopted big data to describe tourism phenomena or tourists' behavior. Merely explaining phenomena is of limited benefit to the development of tourism theory. In addition to descriptions, further exploration of the internal mechanisms and theories behind focal phenomena should not be neglected. Revealing associated mechanisms and principles can greatly enrich tourism management theories and promote theoretical evolution in the Chinese context.

References

Bao, J., Chen, G., & Jin, X. (2018). China tourism research: A review of publications from four top international journals. *Journal of China Tourism Research*, *14*(1), 1-19. https://doi.org/https://doi.org/10.1080/19388160.2018.1434094

Bi, J. W., Liu, Y., & Li, H. (2020). Daily tourism volume forecasting for tourist attractions. *Annals of Tourism Research*, *83*, 102923. https://doi.org/10.1016/j.annals.2020.102923

Chen, W., Xu, Z. Y., Zheng, X. Y., & Luo, Y. L. (2019). Geo-Tagged Photo Metadata Processing Method for Beijing Inbound Tourism Flow. *Isprs International Journal of Geo-Information*, *8*(12), 556. https://doi.org/10.3390/ijgi8120556

Deng, N., & Liu, J. (2021). Where did you take those photos? Tourists' preference clustering based on facial and background recognition. *Journal of Destination Marketing & Management*, *21*, 100632. https://doi.org/https://doi.org/10.1016/j.jdmm.2021.100632

Deng, N., Liu, J. Y., Dai, Y., & Li, H. (2019). Different cultures, different photos: A comparison of Shanghai's pictorial destination image between East and West. *Tourism Management Perspectives*, *30*, 182-192. https://doi.org/10.1016/j.tmp.2019.02.016

Fan, W., Li, Y., Upreti, B., Liu, Y., Li, H., Fan, W., & Lim, E. (2021). Big Data for Big Insights: Quantifying the Adverse Effect of Air Pollution on the Tourism Industry in China. *Journal of Travel Research*, 004728752110472. https://doi.org/10.1177/00472875211047272

Gawlik, E., Kabaria, H., & Kaur, S. (2011). *Predicting tourism trends with Google Insights*. https://www.researchgate.net/profile/Hardik-Kabaria/publication/265201229_Predicting_tourism_trends_with_Google_Insights/links/54e4a5b00cf276cec171d803/Predicting-tourism-trends-with-Google-Insights.pdf

Hao, J. X., Wang, R., Law, R., & Yu, Y. (2021). How do Mainland Chinese tourists perceive Hong Kong in turbulence? A deep learning approach to sentiment analytics. *International Journal of Tourism Research*, *23*(4), 478-490. https://doi.org/10.1002/jtr.2419

Hu, M., & Song, H. (2019). Data source combination for tourism demand forecasting. *Tourism Economics*, *26*(7), 1248-1265. https://doi.org/10.1177/1354816619872592

Hu, M. M., Xiao, M. Q., & Li, H. Y. (2021). Which search queries are more powerful in tourism demand forecasting: searches via mobile device or PC? *International Journal of Contemporary Hospitality Management*, *33*(6), 2022-2043. https://doi.org/10.1108/ijchm-06-2020-0559

Huang, X. K., Zhang, L. F., & Ding, Y. S. (2017). The Baidu Index: Uses in predicting tourism flows -A case study of the Forbidden City. *Tourism Management*, *58*, 301-306. https://doi.org/10.1016/j.tourman.2016.03.015

Jia, S. S. (2020). Motivation and satisfaction of Chinese and US tourists in restaurants: A cross-cultural text mining of online reviews [Review]. *Tourism Management*, *78*, 104071. https://doi.org/10.1016/j.tourman.2019.104071

Jiang, Q. M., Chan, C. S., Eichelberger, S., Ma, H., & Pikkemaat, B. (2021). Sentiment analysis of online destination image of Hong Kong held by mainland Chinese tourists. *Current Issues in Tourism*, *24*(17), 2501-2522. https://doi.org/10.1080/13683500.2021.1874312

Jin, S. (2020). Literature review of domestic tourism big data research in the past 10 years. *Tourism Overview*, *17*, 27-29.

Law, R., & Cheung, S. (2010). The Perceived Destination Image of Hong Kong as Revealed in the Travel Blogs of Mainland Chinese Tourists. *International Journal of Hospitality & Tourism Administration*, *11*(4), 303-327. https://doi.org/10.1080/15256480.2010.518521

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. https://doi.org/https://doi.org/10.1016/j.annals.2019.01.014
Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature, 595*(7866), 189-196. https://doi.org/https://doi.org/10.1038/s41586-021-03660-7
Lazer, D., & Radford, J. (2017). Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*, *43*. https://doi.org/10.1146/annurev-soc-060116-053457
Leung, D., Law, R., & Lee, H. A. (2011). The Perceived Destination Image of Hong Kong on Ctrip.com. *International Journal of Tourism Research*, *13*(2), 124-140. https://doi.org/10.1002/jtr.803
Leung, R., Vu, H. Q., & Rong, J. (2017). Understanding tourists' photo sharing and visit pattern at non-first tier attractions via geotagged photos. *Information Technology & Tourism*, *17*(1), 55-74. https://doi.org/10.1007/s40558-017-0078-3
Li, C., Ge, P., Liu, Z., & Zheng, W. (2020). Forecasting tourist arrivals using denoising and potential factors. *Annals of Tourism Research*, *83*, 102943. https://doi.org/https://doi.org/10.1016/j.annals.2020.102943
Li, H. Y., Hu, M. M., & Li, G. (2020). Forecasting tourism demand with multisource big data. *Annals of Tourism Research*, *83*, 102912. https://doi.org/10.1016/j.annals.2020.102912
Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301-323.
Li, N., Tung, V., & Law, R. (2017). A fuzzy comprehensive evaluation algorithm for analyzing electronic word-of-mouth. *Asia Pacific Journal of Tourism Research*, *22*(6), 592-603. https://doi.org/10.1080/10941665.2017.1308395
Li, S., Li, G., Law, R., & Paradies, Y. (2020). Racism in tourism reviews. *Tourism Management*, *80*, 104100. https://doi.org/10.1016/j.tourman.2020.104100
Li, X., & Law, R. (2020). Network analysis of big data research in tourism. *Tourism Management Perspectives*, *33*, 100608. https://doi.org/https://doi.org/10.1016/j.tmp.2019.100608
Li, X., Law, R., Xie, G., & Wang, S. (2021). Review of tourism forecasting research with internet data. *Tourism Management*, *83*, 104245. https://doi.org/10.1016/j.tourman.2020.104245
Li, X., Li, H. Y., Pan, B., & Law, R. (2021). Machine Learning in Internet Search Query Selection for Tourism Forecasting. *Journal of Travel Research*, *60*(6), 1213-1231. https://doi.org/10.1177/0047287520934871
Li, X., Pan, B., Law, R., & Xiankai, H. (2017). Forecasting tourism demand with composite search index. *Tourism Management*, *59*, 57-66. https://doi.org/10.1016/j.tourman.2016.07.005
Liu, J., Yang, L. Y., Zhou, H. Y., & Wang, S. H. (2021). Impact of climate change on hiking: quantitative evidence through big data mining. *Current Issues in Tourism*, *24*(21), 3040-3056. https://doi.org/10.1080/13683500.2020.1858037
Liu, Y., Huang, K., Bao, J., & Chen, K. (2019). Listen to the voices from home: An analysis of Chinese tourists’ sentiments regarding Australian destinations. *Tourism Management*, *71*, 337-347. https://doi.org/https://doi.org/10.1016/j.tourman.2018.10.004
Liu, Y., Teichert, T., Rossi, M., Li, H., & Hu, F. (2017). Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-

generated reviews. *Tourism Management*, *59*, 554-563. https://doi.org/https://doi.org/10.1016/j.tourman.2016.08.012

Luo, J. Q., Huang, S. S., & Wang, R. W. (2021). A fine-grained sentiment analysis of online guest reviews of economy hotels in China. *Journal of Hospitality Marketing & Management*, *30*(1), 71-95. https://doi.org/10.1080/19368623.2020.1772163

Lv, H., Shi, S., & Gursoy, D. (2021). A look back and a leap forward: a review and synthesis of big data and artificial intelligence literature in hospitality and tourism. *Journal of Hospitality Marketing & Management*, 1-31. https://doi.org/10.1080/19368623.2021.1937434

Pan, B., & Yang, Y. (2017). Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research*, *56*(7), 957-970. https://doi.org/https://doi.org/10.1177/0047287516669050

Rahmadian, E., Feitosa, D., & Zwitter, A. (2021). A systematic literature review on the use of big data for sustainable tourism. *Current Issues in Tourism*, 1-20. https://doi.org/https://doi.org/10.1080/13683500.2021.1974358

Schuckert, M., Liu, X., & Law, R. (2015). A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently. *International Journal of Hospitality Management*, *48*, 143-149. https://doi.org/https://doi.org/10.1016/j.ijhm.2014.12.007

Shoval, N., & Isaacson, M. (2007). Tracking tourists in the digital age. *Annals of Tourism Research*, *34*(1), 141-159. https://doi.org/https://doi.org/10.1016/j.annals.2006.07.007

Statista. (2021). *Coronavirus: economic impact in China*. https://www.statista.com/study/72133/coronavirus-economic-impact-in-china/

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. https://doi.org/10.1016/j.tourman.2018.07.010

Tang, L., Zhang, C. Y., Li, T. F., & Li, L. (2021). A novel BEMD-based method for forecasting tourist volume with search engine data. *Tourism Economics*, *27*(5), 1015-1038. https://doi.org/10.1177/1354816620912995

UNWTO. (2011). *Tourism Towards 2030 / Global Overview*. Retrieved December 10, 2021, from https://www.globalwellnesssummit.com/wp-content/uploads/Industry-Research/Global/2011_UNWTO_Tourism_Towards_2030.pdf

UNWTO. (2020). *TOURISM STATISTICS DATA*. Retrieved December 10, 2021, from https://www.unwto.org/tourism-statistics-data

Vu, H. Q., Li, G., Law, R., & Ye, B. H. B. (2015). Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tourism Management*, *46*, 222-232. https://doi.org/10.1016/j.tourman.2014.07.003

Vu, H. Q., Luo, J. M., Ye, B. H. B., Li, G., & Law, R. (2018). Evaluating museum visitor experiences based on user-generated travel photos. *Journal of Travel & Tourism Marketing*, *35*(4), 493-506. https://doi.org/10.1080/10548408.2017.1363684

Wang, R., Hao, J. X., Law, R., & Wang, J. (2019). Examining destination images from travel blogs: a big data analytical approach using latent Dirichlet allocation. *Asia Pacific Journal of Tourism Research*, *24*(11), 1092-1107. https://doi.org/10.1080/10941665.2019.1665558

Wen, J., & Wu, M. (2020). How special is special interest tourism–and how special are special interest tourists? A perspective article in a Chinese context. *Current Issues in Tourism*, *23*(16), 1968-1972. https://doi.org/https://doi.org/10.1080/13683500.2020.1750575

WTCF. (2021). *The Report on World Tourism Economy Trends (2021)*. https://en.wtcf.org.cn/Research/WTCFReports/2021031220363.html

Wu, C. C., & Chen, D. W. (2021). Tourist versus resident movement patterns in open scenic areas: Case study of Confucius Temple Scenic area, Nanjing, China. *International Journal of Tourism Research*, *23*(6), 1163-1175. https://doi.org/10.1002/jtr.2476
Xu, D., Cong, L., & Wall, G. (2020). Visitors' spatio-temporal behavior at a zoo in China. *Asia Pacific Journal of Tourism Research*, *25*(9), 931-947. https://doi.org/https://doi.org/10.1080/10941665.2020.1802311
Xu, H., Ding, P., & Packer, J. (2008). Tourism research in China: Understanding the unique cultural contexts and complexities. *Current Issues in Tourism*, *11*(6), 473-491. https://doi.org/https://doi.org/10.1080/13683500802475737
Xu, L. (2017, December 15). *Big data promotes China's tourism*. China Daily. http://www.chinadaily.com.cn/a/201712/25/WS5a403ff2a31008cf16da32ea.html
Yang, X., Pan, B., Evans, J., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, *46*, 386–397. https://doi.org/10.1016/j.tourman.2014.07.019
Ying, S., Chan, J. H., & Qi, X. (2020). Why are Chinese and North American guests satisfied or dissatisfied with hotels? An application of big data analysis. *International Journal of Contemporary Hospitality Management*, *32*(10), 3249-3269. https://doi.org/10.1108/IJCHM-02-2020-0129
Zhang, B., Li, N., Shi, F., & Law, R. (2020). A deep learning approach for daily tourist flow forecasting with consumer search data. *Asia Pacific Journal of Tourism Research*, *25*(3), 323-339. https://doi.org/https://doi.org/10.1080/10941665.2019.1709876
Zhang, B. R., Huang, X. K., 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. https://doi.org/10.1080/10941665.2016.1232742
Zhang, K., Chen, Y., & Li, C. L. (2019). Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: The case of Beijing. *Tourism Management*, *75*, 595-608. https://doi.org/10.1016/j.tourman.2019.07.002
Zhang, X. W., Yang, Y., Zhang, Y., & Zhang, Z. L. (2020). Designing tourist experiences amidst air pollution: A spatial analytical approach using social media. *Annals of Tourism Research*, *84*, 102999. https://doi.org/10.1016/j.annals.2020.102999
Zhao, X., Lu, X. N., Liu, Y. Y., Lin, J., & An, J. (2018). Tourist movement patterns understanding from the perspective of travel party size using mobile tracking data: A case study of Xi'an, China. *Tourism Management*, *69*, 368-383. https://doi.org/10.1016/j.tourman.2018.06.026
Zheng, W., Huang, X., & li, Y. (2017). Understanding the tourist mobility using GPS: Where is the next place? *Tourism Management*, *59*, 267-280. https://doi.org/10.1016/j.tourman.2016.08.009
Zheng, W., Zhou, R., Zhang, Z., Zhong, Y., Wang, S., Wei, Z., & Ji, H. (2018). Understanding the tourist mobility using GPS: How similar are the tourists? *Tourism Management*, *71*, 54-66. https://doi.org/10.1016/j.tourman.2018.09.019
Zhong, L., Sun, S., & Law, R. (2019). Movement patterns of tourists. *Tourism Management*, *75*, 318-322. https://doi.org/https://doi.org/10.1016/j.tourman.2019.05.015