Big Data and Analytics in Hospitality and Tourism:

A Systematic Literature Review

Marcello M. Mariani

Henley Business School
University of Reading, Reading, UK
m.mariani@henley.ac.uk

Rodolfo Baggio

Master in Economics and Tourism and
Dondena Center for Research on Social Dynamics
and Public Policy
Bocconi University, Milan, Italy
and
Tomsk Polytechnic University
Tomsk, Russia
rodolfo.baggio@unibocconi.it

International Journal of Contemporary Hospitality Management, 2021
doi: 10.1108/IJCHM-03-2021-0301
[forthcoming]

Abstract

Purpose – This work surveys the body of research revolving around big data and analytics in hospitality and tourism by detecting macro topical areas, research streams and gaps, and develops an agenda for future research.

Design/methodology/approach – This research is based on a systematic literature review of academic papers indexed in the Scopus and Web of Science databases published up to 31 December 2020. The outputs were analyzed using bibliometric techniques, network analysis and topic modeling.

Findings – The number of scientific outputs in research with hospitality and tourism settings has been expanding over the period 2015–2020, with a substantial stability of the areas examined. The vast majority are published in academic journals where the main reference area is neither hospitality nor tourism. The body of research is rather fragmented and studies on relevant aspects, such as big data analytics capabilities, are virtually missing. Most of the outputs are empirical. Moreover, many of the articles collected relatively small quantities of records and, regardless of the time period considered, only a handful of articles mix a number of different techniques.

Research limitations/implications – This study is centered on academic outputs published to the end of 2020 (the last year for which we have full-year data available). Implications are discussed.

Originality/value – This work sheds new light on the emergence of a body of research at the intersection of hospitality and tourism management and data science. It enriches and complements extant literature reviews on big data and analytics, combining these two interconnected topics.

Keywords: Big data, analytics, systematic literature review, Latent Dirichlet Allocation, hospitality, tourism.

Paper type: Literature review.

1 Introduction

The confluence of digital technologies development and government plans and policies strengthening industrial competitiveness, have brought about the digital transformation of human activities and powered the Fourth Industrial Revolution, a socio-economic phenomenon that has profoundly modified the interactions between social and economic actors. This phenomenon, also known as

Industry 4.0 (Rüßmann *et al*., 2015), displays nine underpinning technologies, including big data and analytics. Big data and analytics, in their turn, have been recognized by scholars and media as one of the most relevant technologies (Erl *et al*., 2015) to create value in a world where large amounts of digital data has become the new oil of the digital economy (*The Economist*, 2017). From a methodological point of view, BD-based approaches allow researchers to overcome the liability and difficulties inherent in working with small samples as the entire population under scrutiny can be considered (George *et al*., 2014; Gerard *et al*., 2016).

The term big data (BD) has been deployed since the 1990s and, while some scholars credit its popularization to the US computer scientist and entrepreneur John Mashey, the term was originally used in the computer science field in relation to visualization techniques (Cox and Ellsworth, 1997). In the management and business field, the term BD has been popularized and defined by Gartner analyst Doug Laney at the beginning of this century (Laney, 2001). Since its appearance (Cox and Ellsworth, 1997), BD has been extensively analyzed in scholarly literature across multiple fields, such as information management, supply chain management, marketing, and financial management. However, several scholars have argued that BD is not enough as large volumes of data are not sufficient *per se* to guarantee the generation of relevant knowledge. Indeed, to create value there needs to be analytics defined as a holistic process to access, store, analyze and interpret data for the identification of meaningful patterns in the data.

Over the last decade, and especially during the last five years, the fields of hospitality and tourism have witnessed an increasing use of (and attention to) BD and analytics, with an increasing amount of research produced along these two lines of research. While there have been a couple of literature reviews covering BD in hospitality and tourism up to 2017 (Li *et al*., 2018; Mariani *et al*., 2018), to our knowledge no updated study (covering entirely the 2010–2020 decade) has been produced to understand the extent research lines have evolved over time. As our research will show, these research streams have been expanding considerably over the last four years and, therefore, there is a need to reassess the literature to understand what has been added and what is still unknown. To address this research gap, we systematically review the literature on BD and analytics in the fields of hospitality and tourism, and adopt a quantitative methodology for the analysis, including bibliometric techniques and topic modeling. More specifically, the purpose of this work is to survey and scrutinize the body of research revolving around BD and analytics in hospitality and tourism settings published to 2020, by detecting macro topical areas, research streams and gaps, and developing an agenda for future research. This literature review is not exclusively confined to hospitality and tourism academic journals, but it also provides specific insights on hospitality and tourism academic journals. Compared to existing literature reviews on the topic (Li *et al*., 2018; Mariani *et al*., 2018), our work is distinctively different in three ways. First, while previous work has reviewed articles published to 2017, our intellectual effort is to assess the state of the art of the knowledge in the focal field by considering the most recent three full years (2018–2020) more comprehensively – in addition to the scientific outputs produced before 2018 – and therefore capturing the most recent evolutions. Second, while the two previous literature reviews have focused only on BD and have largely neglected analytics, this study performs queries related explicitly to both BD and analytics, as they are clearly interconnected areas of research and constitute, conjointly, a specific technology underpinning Industry 4.0 (Rüßmann *et al*., 2015). Third, despite describing sophisticated data science techniques, none of the previous review articles have applied these; in this work, we deploy more advanced bibliometric techniques (such as network science-based bibliographic coupling and topic modeling) to gain insights on the literature analyzed. The questions we seek to answer are:

RQ1: What are the most recurrent macro topical areas, research streams and gaps in the literature on BD and analytics pertaining to hospitality and tourism settings?

RQ2: Has there been any evolution of the aforementioned macro topical areas and research streams in the literature since 2017?

To make these contributions, our review study is structured as follows: we first review the relevant literature in the field of BD and analytics, then we illustrate the research methodology deployed to perform the review and the techniques used. Our findings also entail a discussion of articles in relation to research topics, sources of data, type and size of data, data collection methods, analysis, and reporting/visualization techniques, with a focus to the 40 most cited articles across two databases (Scopus and Web of Science) published in hospitality and tourism journals. Finally, we recognize theoretical and methodological knowledge gaps in hospitality and tourism research, draw our conclusions, elucidate the limitations and identify a research agenda for BD and analytics in hospitality and tourism.

2 Big data and analytics

2.1 Big data

The consolidation of internet and digital platforms, and the increasing adoption of smart devices and the Internet of Things (IoT) sensors, have paved the way for an increasing production of data. This data has a different nature and comes from different sources and in different forms. First, while data in the past was mostly recorded or stored in the form of analog data, today data is, increasingly, in the form of digital information which is stored using the binary system, i.e., a series of ones and zeros. Second, data comes from different sources, including devices (mobile roaming data, GPS data, Wi-Fi data), user data (user-generated content, such as text and pictures shared on the internet and social media) and operations (web search data and online booking data) (Li *et al*., 2018). Third, data can be structured (e.g., numbers) or unstructured (e.g., photos).

In the scholarly sphere, the big data notion has emerged since the 1990s, mostly in the context of computer science, when it was used as a term indicating advanced visualization tools and techniques (e.g., Cox and Ellsworth, 1997; Bryson *et al*., 1999) and, later, as a means of storing large quantities of data. It was later popularized to a wider audience at the beginning of the 2000s by Gartner analyst Doug Laney, who detected three major features and characteristics of BD – the 3Vs of volume, velocity and variety (Laney, 2001). Since Laney's definition of BD, there have been many variations on the "V" theme as scholars attempt to define BD (Fosso Wamba *et al*., 2015; Mariani *et al*., 2018; SAS Insights, 2017).

BD has enjoyed increasing scholarly attention over the last decade, within the social sciences in general and management in particular, as researchers have started analyzing the benefits and challenges brought about by BD for research beyond the initial hype (Gandomi and Haider, 2015). For instance, Gerard *et al*. (2016) have emphasized that management research can benefit from BD as it allows management scholars not only to better address existing research questions, but also to develop new research questions and innovative research designs, possibly allowing scholars to achieve better generalizations of their findings. However, while BD can be potentially conducive to better decision-making and performance (Davenport, 2014), and a better interpretation of social phenomena and the ongoing Fourth Industrial Revolution, it also poses challenges in terms of security, privacy and ethical issues (Acquisti *et al*., 2016). Moreover, BD does not automatically translate into better decision-making and performance, as is described in the next subsection.

2.2 Data analytics and big data analytics

Large volumes of multifaceted data are not sufficient to guarantee the generation of relevant knowledge. Indeed, to create value *data analytics* need to be defined as a holistic process to access, store, analyze and interpret data for the identification of meaningful configurations (Fosso Wamba *et al*., 2020). Analytics are a means through which analysts can discover complex patterns of relationships within (large amounts of) data. Data analytics have been broadly classified into four categories that emerge by crossing two dimensions: time (past/present vs. future) and type of knowledge/intelligence created (data/information vs. knowledge). *Descriptive analytics* address the questions "What happened? What is happening?" and aim to generate information about the past and present. They rely on descriptive statistical measures. *Exploratory analytics* address the questions "Why did this happen? Why is this happening?" and seek to generate knowledge about the past and the present. They rely on techniques such as cluster and factor analysis. *Predictive analytics* address the question "What will happen?" and aim to infer information about the future. They rely on techniques such as regression analysis and forecasting techniques. *Prescriptive analytics* address the questions "How to optimize?" and are conducive to generating knowledge about the future. They are based on optimization techniques and experiments.

Several scholars have clearly identified the benefits of analytics in general, and BD analytics (BDA) in particular. More specifically, BDA in business entails the enhancement of business intelligence, price optimization, product positioning, improvement of customer satisfaction, inventory optimization, supply chain risk management, operation streamlining and discovery of business opportunities (Chugh and Grandhi, 2013; Davenport, 2017; Liebowitz, 2013; McAfee *et al*., 2012). In general, descriptive and exploratory analytics aim to increase efficiency, improve processes, and exploit knowledge. Predictive and prescriptive analytics instead support innovation, process re-engineering and knowledge exploration. Data analytics, regardless of the category they belong to, improve organizational performance and agility (Nam *et al*., 2019; Fosso Wamba *et al*., 2015), as well as innovation performance (Kakatkar *et al*., 2020), especially when external data (e.g., data produced by prospective customers on social media) are matched with internal data (e.g., transaction data) (Coker, 2014). In this regard, BDA can be thought as tightly related to business intelligence (Mariani *et al*., 2018).

However, extant literature has found that data analytics and BDA should be matched with appropriate data analytics capabilities. These have been defined as organizational capabilities consisting of tangible (e.g., technology and data), intangible (e.g., data-driven culture), and human resources (e.g., managerial skills and technical skills) (Gupta and George, 2016). For instance, BD is often in unstructured or semi-structured forms, which poses a unique challenge for consumption and analysis (SAS Insights, 2017) that can be overcome through data analytics capabilities. More recently, BDA capabilities (BDAC) have been described as a firm's ability to use talent and technology to retrieve, store and analyze data towards the generation of insight (Mikalef *et al*., 2020). BDAC are contributing to better strategic and operational decisions, higher levels of performance of firms (Ferraris *et al*., 2019; Fosso Wamba *et al*., 2017; Mikalef *et al*., 2020, Rialti *et al*., 2019), and better supply chain management (Fosso Wamba *et al*., 2020, Srinivasan and Swink, 2018).

2.3 Big data and analytics in hospitality and tourism

Hospitality and tourism (H&T) are a context where enormous amounts of data are produced by both H&T service providers and customers. Indeed, tourism firms, destination managers and consumers generate and use large volumes of data and use data analytics to improve decision-making at all levels (Mariani, 2019). For example, user-generated content (UGC) data can be used by both researchers and practitioners to understand tourists', residents' and hospitality service consumers' perceptions and behaviors (e.g., Cheng and Jin, 2019; Ranjbari *et al*., 2020; Zhang *et al*., 2019). Moreover, GPS location data, matched with social media data in traveler smartphones, offer tourism firms insights on what travelers like, or on their needs, thus allowing marketers to profile customers (Dursun and Caber, 2016) and create location and context-specific offers based on travelers' preferences, tastes, needs and behaviors (Buhalis and Sinarta, 2019), which are tracked dynamically (Stylos *et al*., 2021). Web traffic data produced on Destination Marketing Organization's website can be deployed to forecast hotel demand in a tourism destination (Yang *et al*., 2014) or understand which direct flight route to open (Park and Pan, 2018), while web traffic data generated in search engines can help predict tourism demand for a tourism destination (Li *et al*., 2017). Interestingly, BD originating from website traffic, search engine queries, and/or weather information can be deployed or combined to forecast arrivals and hotel occupancy (Pan and Yang, 2017; Sun *et al*., 2019), thus providing important managerial insights for destination marketers and hotel managers.

Recent systematic literature reviews (Li *et al*., 2018; Mariani *et al*., 2018) have highlighted several key aspects of BD research in hospitality and tourism. Li *et al*. (2018) examined in detail the nature of data and found that UGC data is the dominant type of data in tourism research (accounting for 47%), followed by device data (36%) and transaction data (17%). They also underlined that there are key challenges concerning data quality (for instance, quality reliability in online reviews is an issue and biases in Google Trends data), cost (high expenses for sensor devices) and privacy concerns (many tourism stakeholders are not willing to share data) that could be overcome through cooperative academia-industry collaborations. Areas were detected where BD was applied, including tourism demand forecast, sentiment analysis, behavior analysis and tourism recommendation. Furthermore, a prevalence of techniques entailing traditional econometrics techniques was observed. In terms of future directions, it was suggested that research using device data and transaction data should be further expanded despite cost and privacy concerns, and that researchers should use cross-domain and multi-type data to capture the characteristics of the complex tourism system. Moreover, the authors (Li *et al*., 2018) suggested that BD should be used to shed more light on areas such as tourism precaution and crisis management, online marketing, scene spots programming, tourism product design and carrying capacity estimation. Last, they suggested widening the set of data collection and analysis techniques, including trajectory indexing, outlier detection, speech analysis, hybrid techniques, as well as machine learning and deep learning. In a different and independent review, Mariani *et al*. (2018) found that while there was a growth in hospitality and tourism management works that apply analytical techniques to large quantities of data, the research field appeared quite fragmented in scope and rather limited as far as methods and techniques are concerned. The authors observed that most of the BD studies addressed specific research questions in a somewhat isolated way, thus undermining the capability to generate a consistent research stream. They also observed the lack of a conceptual framework that could help identify critical business problems and linking domains such as BD and business intelligence to tourism and hospitality management, and that there were epistemological dilemmas that have not been solved yet. Despite their important contributions, both the existing literature reviews on BD (Li *et al*., 2018; Mariani *et al*., 2018) display several limitations. First, while Li *et al*., (2018) cover scientific works published in the first part of 2017, Mariani *et al*. (2018) only cover works published until 2016. Second, in both the existing literature reviews, the focus is narrowly on BD, and analytics are only briefly and tangentially mentioned. Third, none of the articles deploy data science analytical techniques and advanced bibliometric techniques to make sense of the focal body of knowledge.

3 Methodology

3.1 Research design and data collection

To gain an understanding of the extent to which BD and analytics feature in the hospitality and tourism literature, we performed a systematic quantitative literature review (SQLR) of academic articles indexed in the two major academic databases: Scopus and Web of Science. The method of SQLR was embraced as it has several advantages, including objectivity and replicability (see Tranfeld *et al*., 2003). It has been widely adopted in the social sciences, in the hospitality and tourism domains (e.g., Law *et al*., 2016), and more specifically in assessing hospitality and tourism research revolving around BD (see Li *et al*., 2018; Mariani *et al*., 2018). The approach is particularly suitable to help understand where there is a presence or absence of research in a specific topical area.

In terms of sources, we used the Scopus and Web of Science (WoS) databases as they are the leading sources of indexed academic work in social sciences (Vieira and Gomes, 2009). We selected these databases over other sources (such as Google Scholar) for three reasons. First, Scopus and WoS index most of the scientific production written in English, and Spearman correlations of citation counts between Google Scholar and WoS/Scopus are strong across all subjects (Martin-Martin *et al*., 2018). Second, the combined coverage of the two databases is suitable for this type of literature review (Waltman, 2016). Third, differently from WoS and Scopus, Google Scholar does not provide any user application programming interface (API) to collect documents and conduct bibliometric analyses. Moreover, Google's policy does not allow automatic downloads. Finally, Google Scholar includes everything that can be found via a computerized process (crawling), which means that there is no quality control evaluation on the publication outlets – this makes the content gathered through Scopus and WoS superior (in terms of quality and scientific reliability) to the content gathered through Google Scholar (Halevi *et al*., 2017; Zupic and Čater, 2015).

To summarize, Scopus and WoS were chosen as they allowed us to achieve a very good data coverage, and improved data quality, retrieval and cleanliness. Moreover, to make our analysis stronger, we carried out our bibliometric analyses on both databases separately and compared them as a further robustness check.

Several search criteria were deployed to retrieve the articles. First, in line with Mariani *et al*. (2020) we developed multiple search queries entailing a combination of the focal keywords "big data" and "analytics" with the hospitality and tourism words "*travel*", "touris*", "hospitality", "hotel", "leisure"* in the text, abstract and keywords of the academic outputs. Second, only articles and articles in press were included (conference papers and book chapters were excluded). Third, the retrieved documents had to be written in English. Fourth, as the data used for this study was collected between July 2020 and June 2021, the search was conducted from the beginning of the coverage of both databases up to the 31 December 2020. After eliminating duplicate records and articles which were not directly related to the topic of the analysis, the final dataset used for the analyses contains 883 papers for Scopus and 1,419 for WoS. These papers cover all the BD and analytics studies pertaining to H&T settings, published and indexed over the period 1980–2020.

3.2 Methods and techniques

For the analysis of the literature collected, we employed a set of quantitative methods ranging from bibliometrics – which is increasingly used in H&T research (e.g., Ali *et al*., 2019) – to network analysis, automated text analysis and, in particular, topic modeling. Network analysis is at the basis of most of the bibliometric techniques, such as co-citation and bibliographic coupling used in systematic literature reviews in the management field (Zupic and Čater, 2015). Indeed, network analysis is the reference technique adopted in bibliographic coupling to represent the structure of a scientific field (Small, 1999), and bibliographic coupling networks, citation networks, co-citation networks, topical networks, co-authorship networks and co-word networks have been shown to relate to each other (Yan and Ding, 2012). Generally, network analytical methods provide indicators that can support the assessment of the impact the various contributions have to a field across time and allow analyzing collaborations, highlight the importance of certain issues, recognize influential variables, identify potential research problems or gaps, and draw attention to the boundaries of knowledge within the domain examined. The mapping to a network is done by exposing the relationships (the edges or links) existing among authors, papers and content elements (the vertices or nodes). These methods have been widely used in similar contexts and have proved to be a good complement to narrative approaches to literature reviews based on qualitative content examination (Benckendorff and Zehrer, 2013; Newman, 2004). More specifically, for our study we build the following networks:

- *co-authorship (AUT)*: nodes are the authors and a link between two authors is set when they co-author a paper. The connections are weighted by the number of co-authored papers;
- *cross-citation (CIT)*: nodes are the papers and cross-citations between them are the links weighted by the number of cross-citations; and
- *bibliographic coupling (BIB)*: papers (nodes) are linked when they have at least a common reference. The network is weighted by the number of cross-citations.

The networks were built with the help of VOSviewer (Van Eck and Waltman, 2010), a program for creating, visualizing and exploring bibliometric maps of scientific literature. Input to the program was the list of papers collected and cleaned, as discussed previously.

Among the wealth of metrics provided by network analysis and analysts (Barabási, 2016; da Fontoura Costa *et al*., 2007), we focused on:

- *connectivity*: measured by a fragmentation index as defined by Borgatti (2006):

$$F = 1 - \frac{\sum_k n_k(n_k - 1)}{N(N-1)}$$

  where n_k = number of nodes for each of the k components existing, N = total number of nodes of the network). F is naturally normalized, thus varying from *0* (no fragmentation, the network is entirely connected) to *1* (complete fragmentation, all nodes are isolated); and
- *modularity*: signals the presence of dense subgroups in the network. These are interpretable as communities of interest (for the authors), or themes on which research has focused (papers coupling and cross-citations). It is evaluated with a stochastic algorithm and measured by a normalized modularity index Q.

A community detection algorithm identifies groups of nodes (communities, modules or clusters) that are more connected among them than to other nodes in the network. The modularity index Q measures strength of the division into communities and is roughly the ratio between the number of nodes inside the communities and that one would expect from a purely random arrangement. Q is formally defined as:

$$Q = \sum_k (e_{kk} - a_k^2)$$

where a_k is the total number of links in module k and e_{kk} is the expected value of the same quantity in a network with the same communities, but having links distributed randomly with respect to those communities. Q is normalized (i.e.: $0 \leq Q \leq 1$), $Q = 0$ means no detectable subdivision, $Q = 1$ complete subdivision (the modules are completely disconnected from one another). The modularity algorithm used is the one proposed by Traag *et al.* (2019). The algorithm is recursive and iteratively assigns the nodes to different clusters checking the value of the modularity index. The algorithm stops when Q is at its maximum value.

For a better understanding of the main themes present in our collection we then employed a data-driven approach using text analytic methods. We built a corpus of documents, each containing title, abstract and keywords of the different papers. The corpus was pre-processed by recognizing the different words (tokenization), removing punctuation and common terms (stop-words), and normalizing and standardizing the terms contained in the corpus via a lemmatization (transformation of all inflected words into their basic forms) (Anandarajan *et al.*, 2019). These 'cleaned' documents were analyzed firstly with traditional statistical techniques to derive the distribution of the most frequently used words and 2-grams (contiguous sequence of two words).

The topic modeling method chosen is the Latent Dirichlet Allocation (LDA), recognized to be the most effective among the many possibilities existing (Jelodar *et al.*, 2019). LDA is a generative, probabilistic model that assumes that each item (document) in a corpus contains a mix of topics made of a certain number of words. The model backtracks to find the set of topics that are likely to have generated the corpus. The outcome is a certain number of topics, each characterized by a probability of being present in the corpus, split out into a number of words that are assigned a probability to belong to that topic.

As for other 'clustering' methods, LDA needs to have the number of topics to consider in advance. Since this is generally not known, it is common practice to perform several trials by asking for different numbers of topics and choose the best possible choice, with the help of some metric, in order to distinguish between topics that may be statistical artifacts and those that are semantically meaningful. Here we use a *coherence score* that measures each topic by assessing the semantic similarity between highly frequent words in the topic. The average value of these scores provides a way to choose the best (highest coherence) number of topics. In our case, given the high number of topics, we choose to have the system converge to 30 topics (coherence scores are: 0.474 for Scopus and 0.587 for WoS). All calculations were carried out with the Gensim Python package (Řehůřek and Sojka, 2010).

Once the topics were identified, we built a further network in which papers and topics are the nodes, and the links are weighted with a similarity score. The similarity between two papers then assessed checking the probability with which each topic is represented in the document and using a measure based on the Hellinger distance, a commonly used metric to measure the similarity between two probability distributions (P and Q). The distance is calculated as (Deza and Deza, 2016):

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$$

and the similarity measure used is obviously: $H_S = 1-H(P,Q)$. Both measures are normalized (i.e.: $0 \leq H \leq 1$). The network is fully connected and to improve the efficiency of the calculations, we adopted a threshold $H_S = 0.2$ for the link weights. Although relatively arbitrary, the choice seems reasonable and trials with other values did not change the outcomes substantially.

As a final step we consider the subnetwork of the topics. Two topics are linked if they are present in the same paper.

4 Results and discussion

We detect an exponential growth of interest in investigating and using BD and analytics within the hospitality and tourism domain, starting in 2010. This holds also in other areas that are somehow related. Interestingly, the number of scientific outputs revolving around BD and analytics has been expanding by a multiplier of 7 over the period 2015–2020, and by a multiplier of 10 if we consider the cumulative distribution. Figure 1 shows the cumulative distribution of the number of papers for the sample collected from both databases.

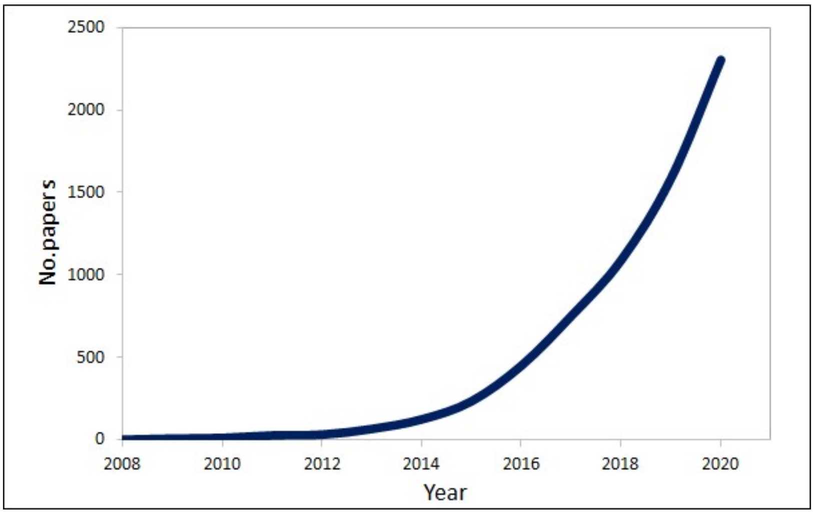

Figure 1 The cumulative distribution of the number of outputs

Besides that, a remarkable observation is that journals mainly focused on tourism and hospitality contain only a low share (27.7% in Scopus and 38.3% in WoS) of the BD and analytics papers, while the great majority (72.3% in Scopus and 61.7% in Wos) belong to outlets typically dedicated to transport, computer science, regional science etc. (e.g., Batty, 2013; Chu *et al.*, 2019; Fan and Gordon, 2014; Hawelka *et al.*, 2014; Liu *et al.*, 2013; Mocanu *et al.*, 2013; Preis *et al.*, 2020; Renjith *et al.*, 2020; Sun *et al*, 2016; Wood *et al.*, 2013). As already noted elsewhere (Mariani *et al.*, 2018), this may signal a certain reluctance of the scholars active in the field to use advanced quantitative approaches and computational algorithms, a certain lack of expertise in computer programming languages, and a scarcity of good hardware and software resources (that may be readily available in other academic departments). Table I contains all the measures calculated for the networks analyzed.

Table I The measures for all networks analyzed

		Scopus				WoS			
		Authors	Countries	Papers	Topics	Authors	Countries	Papers	Topics
No. nodes (order)		2243	75	883	30	4300	83	1419	30
No. links (size)		4879	215	19952	87	9085	369	66702	87
No. components		421	10	129	11	907	8	88	10
No. isolated nodes		68	8	125	10	91	7	84	9
Max component order		835	64	752	20	518	76	1329	21
Max component order (%)		37.2%	85.3%	85.2%	66.7%	12.0%	91.6%	93.7%	70.0%
Fragmentation index	(*)	0.860	0.272	0.275	0.563	0.984	0.163	0.123	0.517
Average degree	(*)	0.002	0.077	0.051	0.200	0.001	0.108	0.066	0.200
Average clustering coefficient	(*)	0.765	0.447	0.400	0.488	0.828	0.518	0.439	0.457
Average closeness	(*)	0.031	0.263	0.284	0.419	0.003	0.340	0.369	0.458
Average betweenness	(*)	0.000	0.014	0.001	0.010	0.000	0.012	0.001	0.011
Average eigenvalue centrality	(*)	0.002	0.080	0.019	0.134	0.001	0.080	0.016	0.137
Average local efficiency	(*)	0.774	0.526	0.557	0.537	0.836	0.624	0.632	0.523
Density	(*)	0.002	0.077	0.051	0.200	0.001	0.108	0.066	0.200
Diameter (LCC)	(*)	11	7	8	3	17	4	6	3
Average path length (LCC)	(*)	4.549	2.449	2.480	1.632	6.605	2.131	2.306	1.629
Assortativity	(*)	0.265	-0.249	0.194	-0.293	0.557	-0.325	0.160	-0.344
Global efficiency	(*)	0.035	0.346	0.328	0.312	0.004	0.437	0.419	0.338
Gini index degrees	(*)	0.416	0.556	0.600	0.540	0.366	0.554	0.530	0.523
Modularity (LCC)	(*)	0.781	0.300	0.326	0.089	0.866	0.221	0.368	0.140
No. communities (LCC)	(*)	23	6	6	3	22	6	4	3

LCC = largest connected compoment

The collaboration networks of the authors (co-authorship) and the countries of their affiliations are shown in Figure 2.

The authors' networks are highly fragmented, although a detectable connected component exists in both cases. The high fragmentation and the distribution of the production in the authors' data suggest a rather sporadic approach or involvement into the topics examined (with the exception of a few authors that contributed more papers). As far as the affiliations of the authors are concerned, the networks (Figure 2B) are more compact than those of authors. The distribution of authors and papers is – in line with scientific production in the wider social sciences – uneven and three countries (China, USA and the UK) provide more than half of the total production. The fragmentation of the authors' networks is further confirmed by a modularity analysis of the major connected component that shows, for both sources, a relatively high number of clusters with a good separation assessed by a relatively high modularity index.

The examination of the content of the papers was conducted considering the titles, abstracts, and keywords of the papers. In line with previous studies (e.g., see Mariani *et al.*, 2018), this is sufficient to extrapolate the main topics and issues covered in the literature. Figure 3 shows the word clouds of the most frequent terms (single words and 2-grams) contained in the two corpuses; a list of the 20 most frequent terms is reported in Table .

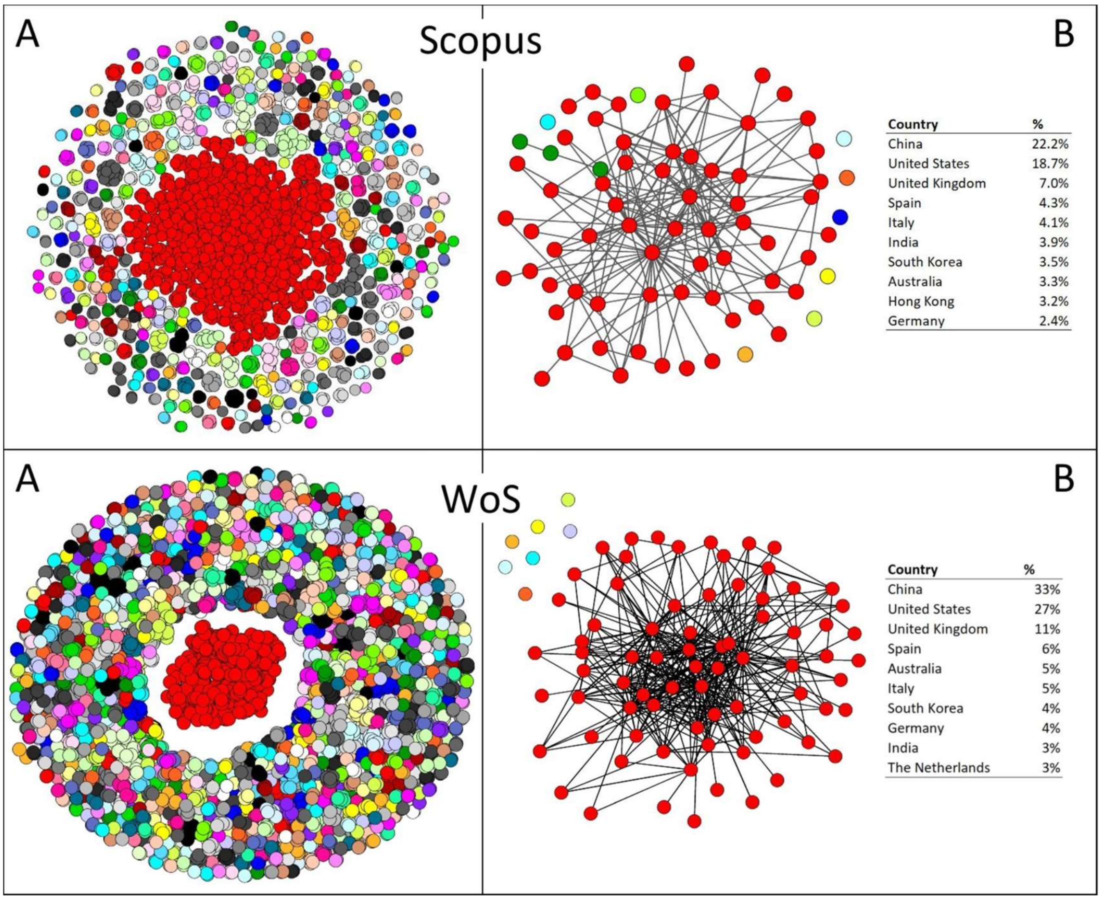

Figure 2 The networks of authors (A), and countries (B), the ten most represented countries are listed

Figure 3 Word cloud of the most common terms used in the corpus

Table II Most frequently used words and 2-grams

	Scopus				WoS				
	Words		2-grams			Words		2-grams	
Rank	Term	Occurr.	Term	Occurr.	Rank	Term	Occurr.	Term	Occurr.
1	data	44.4%	big_data	1.6%	1	data	47.9%	big_data	1.7%
2	tourism	16.8%	social_media	0.5%	2	tourism	20.8%	social_media	0.7%
3	travel	15.0%	data_analytics	0.3%	3	hotel	18.3%	online_reviews	0.5%
4	tourist	9.3%	online_reviews	0.2%	4	online	14.2%	data_analytics	0.3%
5	social	9.0%	machine_learning	0.2%	5	travel	13.5%	machine_learning	0.2%
6	time	8.6%	data_mining	0.2%	6	social	11.5%	user_generated	0.4%
7	analytics	8.3%	travel_behavior	0.2%	7	model	11.5%	customer_satisfaction	0.2%
8	information	8.1%	decision_making	0.2%	8	service	10.5%	sentiment_analysis	0.2%
9	urban	7.4%	real_time	0.1%	9	analytics	8.7%	methodology_approach	0.2%
10	online	6.6%	mobile_phone	0.1%	10	information	8.2%	design_methodology	0.2%
11	model	6.4%	travel_demand	0.1%	11	time	8.1%	decision_making	0.2%
12	management	6.0%	data_analysis	0.1%	12	city	7.6%	online_travel	0.1%
13	media	5.8%	media_analytics	0.1%	13	media	7.4%	data_analysis	0.1%
14	transportation	5.6%	hospitality_industry	0.1%	14	urban	6.9%	data_mining	0.1%
15	hotel	5.3%	smart_tourism	0.1%	15	destination	6.7%	mobile_phone	0.1%
16	destination	5.1%	case_study	0.1%	16	tourist	6.6%	text_mining	0.1%
17	network	5.1%	tourism_industry	0.1%	17	hospitality	6.5%	real_time	0.1%
18	traffic	5.1%	smart_city	0.1%	18	satisfaction	6.2%	smart_tourism	0.1%
19	spatial	5.0%	sentiment_analysis	0.1%	19	customer	6.0%	deep_learning	0.1%
20	systems	4.6%	time_series	0.1%	20	management	6.0%	destination_image	0.1%

Their similarity appears to be high, which is expected given the 'concentration' of the issues discussed. This is confirmed by the cosine similarity of the sets, which is 0.68 for the single words and 0.6 for the 2-grams. The main topics identified by the LDA algorithm are shown in Table .

Table III The top ten topics along with their ten most relevant terms

Scopus

Topic	Terms								
Topic 1	data	travel	predict	model	time	trip	transport	city	system	learn
Topic 2	travel	data	use	transport	big	tourism	model	urban	servic	system
Topic 3	data	online	custom	review	travel	analysis	hotel	analytic	use	predict
Topic 4	data	media	social	analysi	onlin	hotel	model	custom	big	review
Topic 5	data	model	travel	analyt	big	hotel	media	social	transport	system
Topic 6	data	mobil	pattern	urban	tourism	spatial	travel	activ	analysi	social
Topic 7	data	big	analysi	tourism	custom	inform	use	manag	travel	mobil
Topic 8	data	model	review	social	hotel	onlin	analyt	big	use	media
Topic 9	data	model	urban	tourism	analysi	travel	use	tourist	destin	big
Topic 10	data	tourism	tourist	method	big	model	analysi	inform	travel	forecast

WoS

Topic	Terms									
Topic 1	review	onlin	data	model	analysi	use	manag	hotel	satisfact	custom

Topic 2	data	tourism	travel	use	big	tourist	inform	analysi	model	differ
Topic 3	data	mobil	health	time	travel	urban	servic	human	use	big
Topic 4	data	model	research	review	onlin	big	manag	tourism	travel	social
Topic 5	research	hospit	analysi	tourism	review	data	model	industri	use	big
Topic 6	hotel	data	manag	big	experi	onlin	review	analysi	tourism	citi
Topic 7	data	comput	tourism	inform	big	use	system	analysi	develop	fog
Topic 8	data	tourism	model	analysi	travel	big	network	use	servic	research
Topic 9	review	data	onlin	hotel	model	differ	custom	tourist	base	network
Topic 10	onlin	review	hotel	custom	travel	satisfact	servic	differ	attribut	analysi

The coherence score calculated (0.474 for Scopus and 0.587 for WoS) displays a good level and, also in this case, the similarity between the two is quite high (the cosine similarity is 0.94).

The similarity also holds when considering the "evolution" of the issues addressed. If we split the two corpuses into two periods – papers published up to 2017 (included) and after 2017 – the cosine similarity for the two sets is shown in Table . In other words, comparing the most relevant or frequent concepts and ideas discussed in the papers published up to 2017 with those published after 2017, there is practically no difference since the similarity of the terms (words and 2-grams) and topics extracted from the papers of the two periods for both databases are quite high.

Table IV Similarity of words and topics for the two time periods considered (up to 2017 included and after 2017)

	Words	**2-grams**	**Topics**
Scopus	0.77	0.61	0.92
WoS	0.76	0.60	0.86

This similarity in the themes addressed by the papers collected, is also confirmed by the analysis of the bibliographic coupling networks (see numeric values in Table I). They display a relatively compact structure and the modularity analysis of the major component uncovers a small number of communities with a low modularity index, i.e. with a low separation between them. The same holds for the networks of topics that also have large, connected components that display a non-existent separation (see Figure 4).

To further confirm the similarity between all the networks extracted we defined (for each network) a feature vector summarizing their structural properties (i.e., the values marked (*) in Table I). The results reported in Table V speak for themselves – they indicate that all the networks built with the data from the two databases (Scopus and WoS) have a significantly high similarity, thus showing a practically identical structural configuration. This is due to the large overlap that exists between the works included in the search performed (see the lists in Table) and, as is clear from Table VII, the sets of the most cited works are consistent across platforms (Scopus and WoS), with a minor exception – the work by Batty (2013) only appears in the Scopus ranking because the journal *Dialogues in Human Geography* only started being indexed in WoS in 2015, while that article was published in 2013.

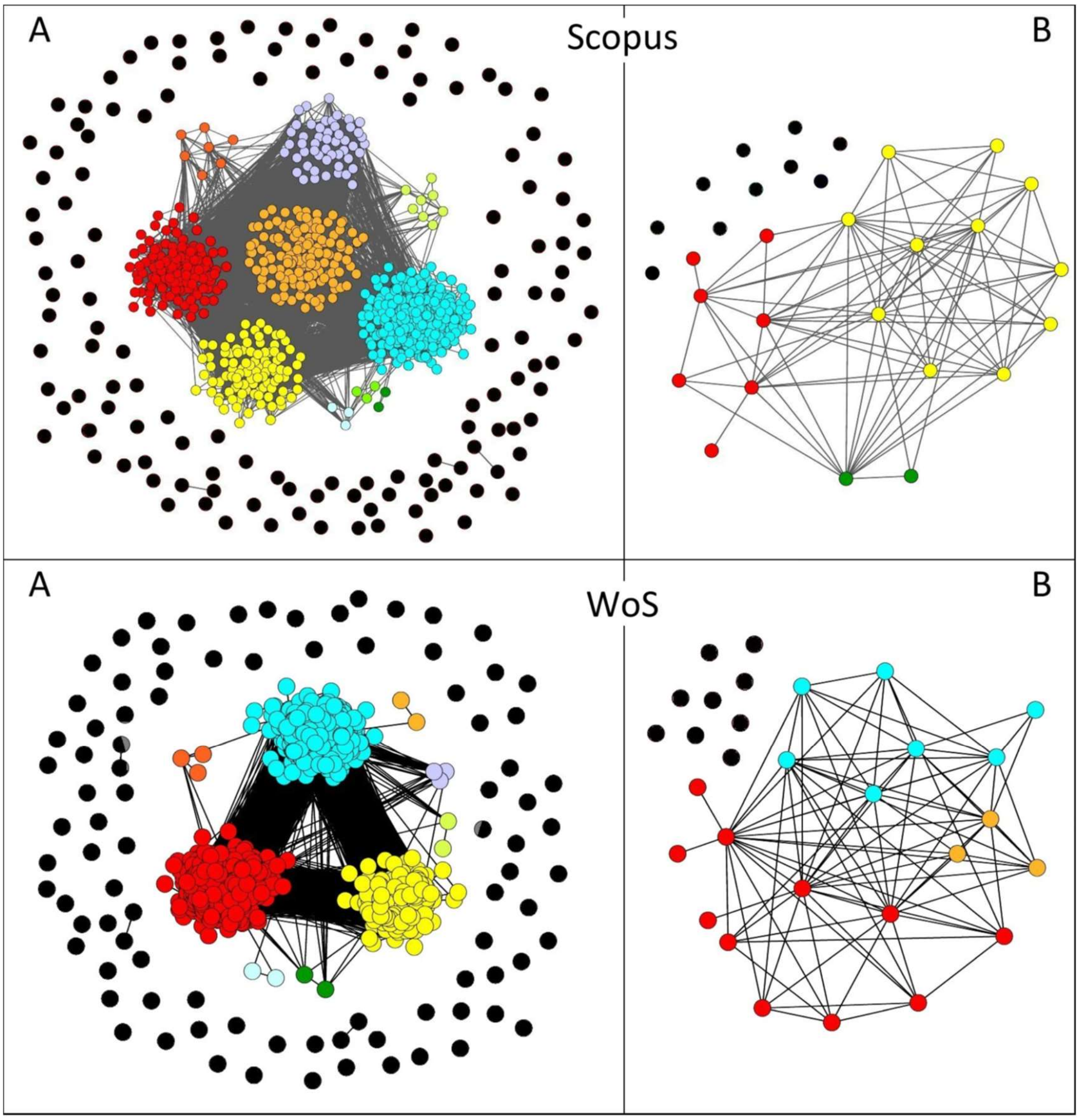

Figure 4 The bibliographic coupling (A) and LDA topics (B) networks. Both show the communities found with a modularity analysis.

Table V The Cosine similarity between all the networks' topological features

Network	Similarity
Authors	0.977
Countries	0.964
Papers	0.995
Topics	0.996

The most represented journals in the entire database of articles (see Table VII) are, not surprisingly, those considered among the top journals in the tourism and hospitality domain: *Tourism Management*, *International Journal of Hospitality Management*, *International Journal of Contemporary Hospitality Management*, *Journal of Travel Research*.

If we the focus on the ten articles with the highest number of citations at the time of data retrieval (see Table), we find that very few have been published in hospitality and tourism journals.

Table VI The top ten papers by citation count

Scopus

Rank	Authors	Title	Journal	Year
1	Gretzel, U. *et al*.	Smart tourism: foundations and developments	*Electronic Markets*	2015
2	Batty, M.	Big data, smart cities and city planning	*Dialogues in Human Geography*	2013
3	Sun, Y. *et al*.	Internet of Things and big data analytics for smart and connected communities	IEEE *Access*	2016
4	Xiang, Z. *et al*.	What can big data and text analytics tell us about hotel guest experience and satisfaction?	*International Journal of Hospitality Management*	2015
5	Wood, S.A. *et al*.	Using social media to quantify nature-based tourism and recreation	*Scientific Reports*	2013
6	Fan, W. and Gordon, M.D.	The power of social media analytics	*Communications of the ACM*	2014
7	Xiang, Z. *et al.*	A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism	*Tourism Management*	2017
8	Guo, Y. *et al*.	Mining meaning from online ratings and reviews: tourist satisfaction analysis using Latent Dirichlet Allocation	*Tourism Management*	2017
9	Zhong, C. *et al*.	Detecting the dynamics of urban structure through spatial network analysis	*International Journal of Geographical Information Science*	2014
10	Toole, J.L. *et al*.	The path most traveled: travel demand estimation using big data resources	*Transportation Research Part C: Emerging Technologies*	2015

WoS

Rank	Authors	Title	Journal	Year
1	Gretzel, U. *et al*.	Smart tourism: foundations and developments	*Electronic Markets*	2015
2	Xiang, Z. *et al*.	What can big data and text analytics tell us about hotel guest experience and satisfaction?	*International Journal of Hospitality Management*	2015
3	Sun, Y.C. *et al*.	Internet of Things and big data analytics for smart and connected communities	IEEE *Access*	2016
4	Wood, S.A. *et al*.	Using social media to quantify nature-based tourism and recreation	*Scientific Reports*	2013
5	Xiang, Z. *et al*.	A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism	*Tourism Management*	2017

6	Guo, Y. *et al*.	Mining meaning from online ratings and reviews: tourist satisfaction analysis using Latent Dirichlet Allocation	*Tourism Management*	2017
7	Fan, W. and Gordon, M.D.	The power of social media analytics	*Communications of the ACM*	2014
8	Zhong, C. *et al*.	Detecting the dynamics of urban structure through spatial network analysis	*International Journal of Geographical Information Science*	2014
9	Toole, J.L. *et al*.	The path most traveled: travel demand estimation using big data resources	*Transportation Research Part C: Emerging Technologies*	2015
10	Chen, C. *et al*.	The promises of big data and small data for travel behavior (aka human mobility) analysis	*Transportation Research Part C: Emerging Technologies*	2016

The most represented journals in the entire database of articles (Table) are, not surprisingly, those considered among the top journals in the tourism and hospitality domain: *Tourism Management, International Journal of Hospitality Management, International Journal of Contemporary Hospitality Management, Journal of Travel Research*.

Table VII Most represented journals for both databases

Scopus		WoS	
Rank	**Journal**	**Rank**	**Journal**
1	*Tourism Management*	1	*Sustainability*
2	*Sustainability*	2	*Tourism Management*
3	*Transportation Research Part C: Emerging Technologies*	3	*International Journal of Hospitality Management*
4	*International Journal of Hospitality Management*	4	*International Journal of Contemporary Hospitality Management*
5	*ISPRS International Journal of Geo-Information*	5	*ISPRS International Journal of Geo-Information*
6	IEEE *Access*	6	*Transportation Research Part C: Emerging Technologies*
7	*International Journal of Contemporary Hospitality Management*	7	IEEE *Access*
8	*IEEE Transactions on Intelligent Transportation Systems*	8	*Journal of Travel Research*
9	*Current Issues in Tourism*	9	*Journal of Hospitality and Tourism Technology*
10	*Transportation Research Record*	10	*Tourism Review*

Table VIII Big data and analytics works in hospitality and tourism (Top 40 most cited work at the time of retrieval in Scopus and WoS, without duplication; in "type of data and size" an asterisk indicates large quantities of data > 100,000 records)

Article (author and title)	**Macro-topical area**	**Research topic**	**Type of paper (conceptual, review, empirical)**	**Sources of data**	**Type of data and size**	**Data collection methods**	**Data analysis techniques**	**Data reporting and visualization**
Alaei *et al*. (2019), Sentiment Analysis in Tourism: capitalizing on big data	Methodological contributions shedding light on a specific technique or family of techniques	Sentiment analysis approaches applied in tourism	Review/methodology	N/A	N/A	N/A	N/A	N/A
Batista e Silva *et al*. (2018), Analyzing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Generates a dataset describing tourist density at high spatial resolution, with monthly breakdown for the European Union	Empirical	EUROSTAT, Booking.com, Tripadvisor	716,000 establishments in Booking.com and Tripadvisor combined	Secondary data from EUROSTAT, and collection from Booking.com and Tripadvisor	Statistical software and geographical information systems (GIS)	Tourist density grids, density maps, tables
Batty (2013), Big data, smart cities and city planning	Knowledge and value creation	Introduces urban big data, explains its importance for cities and short-term thinking and calls for new theory on smart travel card data in Greater London	Conceptual	N/A	N/A	N/A	N/A	N/A
Brandt *et al*. (2017), Social media analytics and value	Knowledge and value creation	Uses Twitter data for San Francisco and kernel density						

creation in urban smart tourism ecosystems		estimation and Latent Dirichlet Allocation to captures spatial patterns within the city	Empirical	600,000 geo-tagged Twitter posts	Structured and unstructured (*)	Data purchased from Twitter	Kernel density estimation and Latent Dirichlet Allocation	Heat maps, figures, tables
Buhalis and Foerste (2015), SoCoMo marketing for travel and tourism: empowering co-creation of value	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	Proposes social context mobile (SoCoMo) marketing as a new framework that enables marketers to increase value for all stakeholders at the destination	Conceptual	N/A	N/A	N/A	N/A	N/A
Buhalis and Leung (2018), Smart hospitality—interconnectivity and interoperability towards an ecosystem	Knowledge and value creation (in smart cities context)	Conceptualizes the smart and agile hospitality enterprises of the future and proposes a smart hospitality ecosystem. The model enables fully integrated applications, using big data to enhance hospitality decision-making	Conceptual	N/A	N/A	N/A	N/A	N/A
Buhalis and Sinarta (2019), Real-time co-creation and nowness service: lessons from tourism and hospitality	Knowledge and value creation (in smart cities context)	Conceptualizes the integration of real-time consumer intelligence, dynamic big data mining, artificial intelligence, and contextualization to illustrate service co-creation	Conceptual	N/A	N/A	N/A	N/A	N/A

Cai et al.(2014), Siting public electric vehicle charging stations in Beijing using big-data informed travel patterns of the taxi fleet	Mapping, identification and representation of tourists, tourists' behaviors, tourist attractions, destinations and trips	Based on trajectory data of more than 11,800 taxis, evaluates how travel patterns mined from big data can inform public charging infrastructure development	Empirical	Trajectory data of 11,880 taxis	Unstructured (*)	Not specified	Probability density distributions	Charts, tables
Chen et al. (2016), The promises of big data and small data for travel behavior (aka human mobility) analysis	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Literature review trying to reconcile two separate fields, both involving understanding and modeling of how individuals move in time and space: "travel behavior analysis" and "human mobility analysis"	Conceptual	N/A	N/A	N/A	N/A	N/A
Cheng and Jin (2019), What do Airbnb users care about? An analysis of online review comments	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	By analyzing a "big data" set of Airbnb online reviews through text mining and sentiment analysis, the study finds that Airbnb users tend to evaluate their experience based on a frame of reference derived from past hotel stays. Key attributes identified in the data include	Empirical	181,263 Airbnb online reviews	Structured and unstructured (*)	Inside Airbnb website	Text mining through unsupervised learning, Leximancer	Charts, figures, tables

		“location”, “amenities” and “host”						
Chua *et al*. (2016), Mapping Cilento: using geotagged social media data to characterize tourist flows in southern Italy	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Analyzes geotagged social media data from Twitter to characterize spatial, temporal and demographic features of tourist flows in Cilento	Empirical	72,031geotagged Tweets	Structured and unstructured	Twitter's stream API, Twitter's REST API	Trajectory mining using graphs	Charts, figures, FlowSampler for flow visualization
Del Vecchio *et al*. (2018), Open innovation and social big data for sustainability: evidence from the tourism industry	Knowledge and value creation	Explores a set of regional tourist experiences related to a Southern European region and destination, to derive patterns and opportunities of value creation generated by big data in tourism	Empirical	7 social media accounts (including Instagram accounts, travel blogs, etc.). Posts, tweets/retweets up to 6,028	Structured and unstructured	Tracking accounts	Keyhole for clustering and Buzztrack for sentiment analysis	Charts, figures, tables
Fan and Gordon (2014), The power of social media analytics	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Discusses the benefits of social media analytics, the underlying stages of the analytics process, the most common social media analytic techniques, and the ways analytics creates business value	Conceptual	N/A	N/A	N/A	N/A	N/A

Fuchs *et al*. (2014), Big data analytics for knowledge generation in tourism destinations – a case from Sweden	Knowledge and value creation	Presents a knowledge infrastructure implemented at the Swedish mountain tourism destination, Åre and examples of use by tourism managers	Empirical	Web search, booking and feedback data (e.g., survey-based, user-generated content)	Structured and unstructured	Data warehouse (DW) including facts and dimensions tables	Online analytical processing (OLAP); support vector machines (SVM), Naïve Bayes (NB) and k-nearest neighbor (KNN)	Html-based web application
Gao *et al*. (2013), Discovering Spatial Interaction Communities from Mobile Phone Data	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	The study adopts an agglomerative clustering algorithm based on a Newman-Girvan modularity metric. By proposing an alternative modularity function incorporating a gravity model, this work helps discover the clustering structures of spatial-interaction communities using mobile phone datasets	Empirical	Phone records of 1,000,000 mobile subscribers in Harbin, China. A total of 74,000,000 phone call records	Structured (*) 74,000,000 phone call detail records	Not disclosed	Spatial and temporal analysis using networks	Tables, maps, figures matched with Google Earth
García-Palomares *et al*. (2015), Identification of tourist hot spots based on social networks: a comparative analysis of European metropolises using photo-sharing services and GIS	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Use of photo-sharing services for identifying and analyzing the main tourist attractions in eight major European cities	Empirical	Panoramio photos	Unstructured (*)	Panoramio website API + ArcGIS	Density graphs, spatial autocorrelation	Standard tables and Anselin Local Moran's I graph

Gretzel *et al*. (2015), Smart tourism: foundations and developments	Knowledge and value creation (in a smart cities context)	Defines smart tourism, sheds light on current smart tourism trends, and lays out its technological and business foundations	Conceptual	N/A	N/A	N/A	N/A	N/A
Guo *et al*. (2017), Mining meaning from online ratings and reviews: tourist satisfaction analysis using Latent Dirichlet Allocation	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	Uses Latent Dirichlet analysis (LDA) to identify key dimensions of customer service in online reviews. LDA uncovers 19 controllable dimensions that are key for hotels to manage their interactions with visitors	Empirical	Tripadvisor online reviews	Structured and unstructured (*)	Web crawler	LDA, correspondence analysis, ANOVA, regression analysis, perceptual mapping	Charts, tables, maps
Hasan and Ukkusuri (2014), Urban activity pattern classification using topic models from online geo-location data	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Develops a model to analyze large-scale geo-location data from social media to infer individual activity patterns	Empirical	Geo-located Twitter check-ins	Structured and unstructured (*)	Web crawler	Topic modeling (LDA)	Charts, tables
Hasan *et al*. (2013), Spatiotemporal patterns of urban human mobility	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Use of smart subway fare card transactions data to model urban mobility patterns. The popularity of places in the city is deployed as an interaction	Empirical	Smart card transactions over a three-month period of public transport users of London	Structured and unstructured	Data was collected by Transport for London (TfL) for operational purposes	Visit probability function	Charts, tables

		parameter between different individuals						
Khalilzadeh and Tasci (2018), Large sample size, significance level, and the effect size: solutions to perils of using big data for academic research	Methodological contributions shedding light on a specific technique or family of techniques (measurement problem)	Informs tourism and hospitality academia of the effect size, measures for the most commonly used statistical tests when dealing with big data	Methodological	Individuals	Structured	Survey	Tests	Tables
Kim *et al*. (2019), Quantifying nature-based tourism in protected areas in developing countries by using social big data	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Illustrates spatial patterns of visitation using 10 years of Flickr geo-tagged photographs	Empirical	Flickr geotagged photos	Structured and unstructured	OpenStreet Map, Python model, NatCap, InVEST.rec reation	Geographically weighted regression	Charts, tables
Kim *et al*. (2017), What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	Applies sentiment analysis to VirtualTourist online reviews of Paris	Empirical	19,835 online reviews from VirtualTourist	Structured and unstructured	Web crawling program developed in Python	Stanford sentiment analysis tool based on JAVA1.7.0_65, . Co-occurrence analysis	Tables, charts

Kirilenko *et al*. (2018), Automated sentiment analysis in tourism: comparison of approaches	Methodological contributions shedding light on a specific technique or family of techniques	Illustrates the suitability of different types of automated classifiers for applications typical in tourism, hospitality, and marketing studies by comparing their performance to that of human assessors/raters	Methodological and empirical	209 surveys of travelers, 332 Tripadvisor reviews, 200 English-language Twitter messages	Structured and unstructured	Online panel survey, Tripadvisor, Twitter	Sentiment analysis using four types of software (SentiStrength, Deeply Moving, two programs using RapidMiner) vs. human raters. Accuracy, precision and recall for software and Cohen's kappa, Kendall's τ, ratio of opposite classification for human raters	Tables, figures
Li *et al*. (2018), Big data in tourism research: a literature review	Knowledge and value creation	Reviews literature on different types of big data in tourism research, distinguishing UGC data, device data and transaction data. Research focuses, data characteristics, analytic techniques, major challenges and further directions are identified	Literature review	N/A	N/A	N/A	N/A	N/A

Li *et al*. (2017), Forecasting tourism demand with composite search index	Demand evaluation and forecast/prediction	Proposes and tests a framework and procedure to create a composite search index adopted in a generalized dynamic factor model (GDFM). This is used to predict tourist volumes to Beijing	Empirical	Baidu search engine trends, data series and tourist volumes	Structured (*)	Search using search engine	Econometric model	Charts, tables
Liu *et al*. (2017), Big data for big insights: investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	Examines the determinants of hotel customer satisfaction by discriminating among customers by language group. Study of hotel customer reviews written by guests speaking 10 different languages	Empirical	412,784 Tripadvisor online reviews of hotels located in five Chinese cities	Structured (*)	Crawler developed in PHP	Language-detection was through MySQL and the textcat package, regression analysis	Charts, tables
Ma *et al*. (2018), Effects of user-provided photos on hotel review helpfulness: an analytical approach with deep leaning	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	Compares deep learning models with other machine learning techniques to examine the effect of user-provided photos on review helpfulness across two social media	Empirical	24,960 Tripadvisor online reviews and 3,064 Yelp online reviews	Structured and unstructured	Web crawler in Python based on APIs	Deep learning model: neural network for sequence encoding, residual network for image representation, feature fusion	Tables and Figures
Mariani *et al*. (2018), Business intelligence and big data in hospitality and tourism: a systematic literature review	Knowledge and value creation	Performs a systematic literature review of business intelligence and big data	Literature Review	N/A	N/A	N/A	N/A	N/A

Mariani and Borghi (2018), Effects of the Booking.com rating system: bringing hotel class into the picture	Methodological contributions shedding light on a specific technique or family of techniques (data quality)	Investigates the effects of the Booking.com rating system on the distribution of hotel ratings for the overall population of hotels located in London over two years	Empirical	1,228,089 Booking online reviews	Structured and unstructured (*)	Web crawler developed in Python	Nonparametric kernel density estimators, parametric and nonparametric tests	Tables and Figures
Mariani *et al*. (2016), Facebook as a destination marketing tool: evidence from Italian regional destination management organizations	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	Explores how Italian regional destination management organizations (DMOs) strategically employ Facebook to promote and market their destinations, and improves on the current metrics for capturing user engagement	Empirical	Overall number of Facebook posts posted on the official Italian regional DMOs' Facebook pages	Structured (*)	Data extractor based on Facebook APIs	Data parser and analyzer, calculating per post statistics	Tables created through data analyzer module. Graphs created through the data visualizer module
Marine-Roig and Anton Clavé (2015), Tourism analytics with massive user-generated content: A case study of Barcelona	Knowledge and value creation	Studying the online image of Barcelona as transmitted via social media through the analysis of more than 100,000 relevant travel blogs and online travel reviews (OTRs) written in English	Empirical	Heterogeneous, including the travel blogs, webpages, travelogues and travel reviews about Barcelona (250,000 pages)	Unstructured (*)	Data was extracted through Offline Explorer Enterprise (OEE)	Pre-processing: web content mining, language detection, user's hometown, cleaning, debugging. Processing: parser settings and categorizations through site content analyzer	Tables created through word count

Miah *et al*. (2017), a big data analytics method for tourist behaviour analysis	Knowledge and value creation	Using a design science research approach, designs and evaluates a big data analytics method to support strategic decision-making in tourism destination management	Empirical and methodological	Geotagged Flickr photos uploaded by tourists	Unstructured (*)	Flickr API	Textual metadata processing (GATE); geographical data clustering (P-DBSCAN); representative photo identification (visual content representation and Kernel density estimation); time series modeling (MAE)	Maps, charts and tables
Mocanu *et al*. (2013), The twitter of babel: mapping world languages through microblogging platforms	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Survey on worldwide linguistic indicators and trends	Empirical	Large-scale dataset of geotagged tweets	Structured and unstructured (*)	Twitter API	Language detection (Google Chromium Compact Language Detector) – geographical analyses	Maps, charts and tables
Paldino *et al*. (2015), Urban magnetism through the lens of geo-tagged photography	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Tastes of individuals, and what attracts them to live in a particular city or spend their vacation there	Empirical	Geo-tagged photos	Structured: metadata from photos (*)	Flickr API	Identification of resident, tourist and unknown, statistical analysis, network analysis (origin/destination)	Maps, charts, tables

Pan and Yang (2017), Forecasting destination weekly hotel occupancy with big data	Demand evaluation and forecast/prediction	Deploys time-series models incorporating tourism big data sources, including search engine queries, website traffic, and weekly weather information, to construct an accurate forecasting model of weekly hotel occupancy	Empirical	Google queries (Google Correlate), weekly session data on website traffic for the Charleston Area Convention and Visitors Bureau, weekly weather information from National Weather Service for North America and Europe. In addition, STR hotel occupancy data	Structured data	Google Correlate, Charleston Area Convention and Visitors Bureau, STR	Time series models (ARMAX, MSDR)	Charts, tables
Park *et al*. (2016), Using Twitter data for cruise tourism marketing and research	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Illustrates social media analytics using Twitter data referring to cruise travel	Empirical	50,414 cruise related Tweets	Structured and unstructured	ScraperWiki, Twitter API	Frequency analysis, network mapping	Tables, figures
Plaza (2011), Google Analytics for measuring website performance	Methodological contributions shedding light on a specific technique or family of techniques (platform issues and features)	Measures the effectiveness of website visit behavior and length of sessions, depending on the traffic source	Empirical	7,561 entries for 1,092 days drawn from Google Analytics	Structured	Excel	Regressions	Figures, tables

Puiu *et al*. (2016), CityPulse: large scale data analytics framework for smart cities	Methodological contributions shedding light on a specific technique or family of techniques (platform issues and features)	Presents a framework named CityPulse, describes its components, and demonstrates how they interact to support easy development of custom-made applications for citizens	Conceptual	Development of an architecture	N/A	N/A	N/A	N/A
Raun *et al*. (2016), Measuring tourism destinations using mobile tracking data	Demand evaluation and forecast/prediction	Measure space-time tracking data to analyze, monitor and compare destinations based on data describing actual visits	Empirical	Anonymized roaming data of foreign mobile phones (406,590 visits by 215,643 different phone IDs)	Structured (*)	From telcom operator	Statistical analyses, ArcGIS for spatial analyses, binary logistic regression	Charts, tables, maps
Rossetti *et al*. (2016), Analyzing user reviews in tourism with topic models	Methodological contributions shedding light on a specific technique or family of techniques	A description of the topic model method with application focus on the tourism domain	Empirical	Yelp Dataset Challenge; Tripadvisor dataset	Structured and unstructured	Yelp’s existing dataset; Tripadvisor, automatically collected by crawler	K-nearest neighbor user based (KNN-UB), k-nearest neighbor item based (KNN-IB), probabilistic matrix factorization (PMF)	Illustrative examples for selected topics related to multi-criteria dimensions

Salas-Olmedo *et al*. (2018), Tourists’ digital footprint in cities: comparing big data sources	Methodological contributions shedding light on a specific technique or family of techniques (platform issues and features)	Analyzes the digital footprint of urban tourists through big data from Panoramio (sightseeing), Foursquare (consumption) and Twitter (being connected – accommodation)	Empirical	307,062 geolocated Panoramio photographs, 234,159 Tweets, of which 20,076 Foursquare check-ins	Structured and unstructured (*)	Panoramio API	Statistical analyses, MongoDB, ArcGIS for spatial analyses, OLS regression, spatial autocorrelation	Charts, tables, maps
Schuckert *et al*. (2016), Insights into suspicious online ratings: direct evidence from Tripadvisor	Methodological contributions shedding light on a specific technique or family of techniques (data quality)	Examines gap between overall rating and individual ratings, as well as the proportion of suspicious ratings	Empirical	41,572 Tripadvisor reviews	Structured and unstructured	Crawler was developed to retrieve Tripadvisor online review data	Java-based program to parse HTML and XML web pages, regression analyses	Charts, tables
Sun *et al*. (2019), Forecasting tourist arrivals with machine learning and internet search index	Demand evaluation and forecast/prediction	Verifies the Granger causality and co-integration relationship between internet search index and tourist arrivals of Beijing	Empirical	Tourist arrivals, search query data from Baidu Index and Google Trends	Structured	Tourist arrivals from Wind Database	Kernel extreme learning machine, Granger causality test	Tables and charts

Sun *et al*. (2016), Internet of Things and big data analytics for smart and connected communities	Knowledge and value creation (in smart cities context)	Integration of Internet of Things (IoT) and big data analytics for smart connected communities	Conceptual and case study	Design of an IoT system personal sensors, open data and participatory sensing to enhance the services in the area of tourism and cultural heritage with a context-aware recommendation system	N/A	N/A	N/A	N/A
Tenkanen et al. (2017), Instagram, Flickr or Twitter: assessing the usability of social media data for visitor monitoring in protected areas	Methodological contributions shedding light on a specific technique or family of techniques (platform issues and features)	Estimation of visitation in protected areas through social media analytics	Empirical	Tourists' posts on Instagram, Flickr or Twitter	Structured and unstructured	APIs	Spearman correlation, Pearson correlation	Tables and charts
Toole *et al*. (2015), The path most traveled: travel demand estimation using big data resources	Demand evaluation and forecast/prediction	Develops a software system to estimate multiple aspects of travel demand using call detail records (CDRs) from mobile phones in conjunction with open- and crowdsourced geospatial data, census records, and surveys	Empirical	Call detail records, geospatial data, census records, surveys	Structured and unstructured (*)	Data from companies and cities, GIS, OpenStreet Map, surveys	Origin – destination matrices, network analysis, other analytics	Charts, tables, maps

Tu *et al*. (2016), Optimizing the locations of electric taxi charging stations: a spatial–temporal demand coverage approach	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Develops a spatial–temporal demand coverage approach for optimizing the placement of electric taxis charging stations in a space–time context. The study also evaluates the carbon emission generated by used electric taxis	Empirical	Taxi GPS data, transportation network, charging stations	Structured and unstructured	GPD, NAVInfo China	Map matching, segmentation, spatial–temporal demand coverage location model	Charts, tables, maps
Tussyadiah and Zach (2017), Identifying salient attributes of peer-to-peer accommodation experience	Methodological contributions shedding light on a specific technique or family of techniques	Explores themes from Airbnb online reviews to explain major service attributes sought by guests. Conducts a lexical analysis	Empirical	41,560 Airbnb reviews	Structured and unstructured	Inside Airbnb website	Lexical analysis, tokenization, word co-occurrence, regressions	Charts, tables, maps

Vilajosana *et al*. (2013), Bootstrapping smart cities through a self-sustainable model based on big data flows	Methodological contributions shedding light on a specific technique or family of techniques	Elaborates a procedure and a framework to make smart cities happen based on big data exploitation through the API stores concept	Conceptual	N/A	N/A	N/A	N/A	N/A
Wood *et al*. (2013), Using social media to quantify nature-based tourism and recreation	Methodological contributions shedding light on a specific technique or family of techniques measurement	Online posted photos are used to estimate visitation rates and travelers' origins, compared to empirical data showing that crowd-sourced information can serve as reliable proxy for empirical visitation rates	Empirical	Empirical datasets that quantified visitation to 836 sites in 31 countries around the world, plus Flickr metadata	Structured (*)	Dataset + Flickr API	Statistical and spatial analyses	Charts, tables, maps
Xiang *et al*. (2017), A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism	Methodological contributions shedding light on a specific technique or family of techniques (platform issues and features)	Three major online review platforms (Tripadvisor, Expedia and Yelp) are examined comparatively through text analytics in relation to hotel population in Manhattan (NYC). Discrepancies in the representation of the hotel industry on the platforms are identified	Empirical	Online reviews from three different platforms (Tripadvisor, Expedia, Yelp)	Structured and unstructured (*)	Web crawlers written in the Python and Java programmin g languages	Topic modeling (LDA), sentiment analysis, ML techniques, regression analyses	Charts, tables, maps

Xiang *et al*. (2015), What can big data and text analytics tell us about hotel guest experience and satisfaction?	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	Text analytics are produced for a large quantity of Expedia consumer reviews to deconstruct hotel guest experience and examine the association with satisfaction ratings. The association between guest experience and satisfaction is empirically proved	Empirical	Expedia online reviews	Structured and unstructured	Automated web crawler	Text analytics process involving data pre-processing, domain identification, statistical association analysis	Charts, tables
Xu *et al*. (2017), Business intelligence in online customer textual reviews: understanding consumer perceptions and influential factors	Tourists' (residents and service providers) perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services	Examines customer satisfaction and dissatisfaction toward attributes of hotel products and services based on online customer textual reviews	Empirical	3,596 Tripadvisor online reviews	Structured and unstructured	Manual collection	Latent semantic analysis, regression analysis	Charts, tables
Yang *et al*. (2014), Predicting hotel demand using destination marketing organization's web traffic data	Demand evaluation and forecast/prediction	Deploys web traffic volume data of a destination marketing organization (DMO) to predict hotel demand for the destination	Empirical	Google Analytics account of the CACVB website, and (STR) hotel demand and occupancy data for the Charleston area	Structured	Google analytics plus standard data	Statistical and time series forecasts	Charts, tables
Zhao *et al*. (2019), Predicting overall customer satisfaction: big data evidence from hotel online textual reviews	Demand evaluation and forecast/prediction	Predicts overall customer satisfaction using the technical attributes of online textual reviews and customers'	Empirical	127,629 Tripadvisor online reviews	Structured and unstructured (*)	Customized scripts in Python	Regression analyses	Charts, tables

		involvement in the review community						
Zhong *et al*. (2014), Detecting the dynamics of urban structure through spatial network analysis	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Develops quantitatively a spatial structure of urban movements by constructing a weighted directed graph from smart card travel records in Singapore. Graph properties are used to obtain a view of travel demand, urban centers, neighborhoods	Empirical	Smart card data	Structured and unstructured (*)	Data provided by the Singapore Land Transport Authority	Network analysis package (R), community detection analysis (MapEquation), spatial analysis tool (ArcGIS)	Charts, tables, maps
Zhou *et al*. (2015), Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform	Mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips	Faces the challenges inherent in spatial analytics and automates the detection of places of interest in multiple cities based on spatial and temporal features of Flickr images from 2007. RHadoop platform is used	Empirical	Flickr Creative Dataset from Yahoo Lab, containing 99.3 million images, 49 million of which were geotagged	Structured and unstructured (*)	Hadoop cluster, Geospatial Data Abstraction Library (GDAL)	Tag calculation, tag classification	Charts, tables, maps

Regarding the two data sources (Scopus and WoS) used in this work, we need to make a few considerations. As the analyses reported here show, there is a substantial similarity between the two databases, both in relation to the "structures" of the relationships between the main elements (authors, papers, countries) and the content of the items examined (titles, keywords and abstracts of the papers selected). Coupled with the overlap existing in terms of journals covered by the two databases, this suggests that the analysis of the literature, at least in relation to the domain of hospitality and tourism, can be limited to only one of the two.

To better understand the main issues discussed in the literature, Table VIII entails a classification of the 40 most cited articles in the sample, performed consistently with one of the two previous literature reviews (Mariani *et al*., 2018).

The macro-topical areas include: 1) the perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services of tourists, residents and service providers (e.g., Gruss et al., 2020; Mehraliyev et al., 2020; Park et al., 2020; Ying et al., 2020); 2) demand evaluation and forecast/prediction (e.g., Höpken et al., 2020; Li et al., 2020; Sánchez-Medina and C-Sánchez, 2020); 3) mapping, identification and representation of tourists, tourist behaviors, tourist attractions, destinations and trips (e.g., Chun et al., 2020; Ma and Kirilenko, 2020; Zhang et al., 2020); 4) knowledge and value creation (Kubo et al., 2020); 5) methodological contributions, shedding light on a specific technique or family of techniques (e.g., Alaei et al., 2019; Chang et al., 2020), measurement problem (e.g., Khalilzadeh and Tasci, 2017), or data quality issue. The aforementioned macro topical areas have not changed (comparing the period before and after 2017), consistent with the quantitative findings illustrated in Tables 5 and 6.

Interestingly, none of the articles of the wider sample of Scopus and WoS articles (N=2,302) focus on BD analytics capabilities or tested relationships between BDA and hospitality firm performance. This is clearly a major research gap and denotes a significant delay of hospitality and tourism management research, compared to other research conducted in the information management and wider management field (e.g., Fosso Wamba *et al*., 2017; Gupta and George, 2016; Mikalef *et al*., 2020). More specifically, in the information management field, Gupta and George (2016) have deployed the resource-based theory of the firm, to single out the three major components of BDACs: 1) intangible resources (encompassing data-driven culture and intensity of organizational learning); 2) tangible resources (including data, technology and basic resources (such as time and investment)); 3) human resources (entailing managerial skills (such as analytics acumen) and technical skills (such as education and training pertaining to data-specific skills)). Later, some empirical studies (e.g., Mikalef *et al*., 2019a) have shown that BDACs allow streamlining of value chains (e.g., Srinivasan and Swink, 2018), support innovation (Mikalef *et al*., 2019b) and enhance firm competitive performance (Mikalef *et al*., 2020). While there have not been major changes in the macro topical areas dealt with by BD and analytics scholars comparing the period before and after 2017, a few sporadic studies are deploying BD to monitor tourism demand (e.g., Gallego and Font, 2020) or supply (Park *et al*., 2020) during the COVID-19 pandemic. Arguably, studies at the intersection between BD, analytics and COVID-19 in travel and tourism are expected to grow, especially from 2021 (which is not covered in our database).

As far as the type of articles are concerned, most are empirical (e.g., Batista e Silva *et al*., 2018; Chen and Jin, 2019; Li *et al*., 2020), a few explore methodological aspects, sometimes in the form of a narrowly focused review (e.g., Alaei *et al*., 2019; Biswas *et al*., 2020; Fu *et al*., 2019; Khalilzadeh and Tasci, 2018), and only a very few are conceptual in nature (e.g., Buhalis and Sinarta, 2019; Gretzel *et al*., 2015). Despite the presence of a sporadic literature review focused on a very specific subtopic of BD (Xu *et al*., 2020), there are only two articles that review the BD literature broadly (Li *et al*., 2018; Mariani *et al*., 2018). However, as mentioned in section 2.3 of the literature review, none of the review articles focus explicitly on the broad and interconnected field of BD and analytics.

In relation to the sources of data, also for the period 2018 to 2020, the dominant type of data is UGC data (e.g., Cheng and Jin, 2019; Liu *et al*., 2019; Ma *et al*., 2020; Park *et al*., 2020; Salas-Olmedo *et al*., 2018), followed by device data (e.g., Buning and Lulla, 2020; Kubo *et al*., 2020) and last, transaction data (e.g. Gallego and Font, 2020; Liu *et al*., 2019; Park and Pan, 2018). This seems in line with the findings of another literature review (Li *et al*., 2018). However, after 2017 we observe an increasing number of studies mixing data from different sources (e.g., Batista e Silva *et al*., 2018; Li *et al*., 2020; Park and Pan, 2018; Salas-Olmedo *et al*., 2018).

Regarding the size of data retrieved and processes, most of the articles collected less than 1 million records (Ma *et al*., 2018; Xiang *et al*., 2017; Zhao *et al*., 2019), some between 1 and 3 million (Mariani and Borghi, 2018), and a very few collected more than 3 million (e.g., Raun *et al*., 2016). The only two studies that collected more than 70 million records (Gao *et al*., 2013; Zhou *et al*., 2015) are analysis of the H&T context published in computer science journals.

As far as data collection methods are concerned, most of the researchers using UGC data develop their own crawlers, typically using programming languages such as Python, Java, PHP (Kim *et al*., 2019; Liu *et al*., 2017; Mariani *et al*., 2016; Xiang *et al*., 2017), with Python slightly becoming the dominant programming language. They also typically leverage on Application Programming Interfaces (API) to retrieve data from the major social media platforms, such as Facebook, Twitter and travel review websites like Tripadvisor (e.g., Chua *et al*., 2016; Ma *et al*., 2018; Salas-Olmedo *et al*., 2018). Researchers using device data typically buy or request data directly from the companies owning the data, such as telecommunication companies (Gao *et al*., 2013; Kubo *et al*., 2020; Park *et al*., 2020; Raun *et al*., 2016). Scholars deploying transaction data either purchase them from the owners, from data analytics companies or other types of commercial providers (e.g., Gallego and Font; Park and Pan, 2018) or retrieve them using crawlers (Liu *et al*., 2018).

For the analytical techniques, pre-2017 studies tended to use a narrow number of non-advanced techniques, such as regression analyses (Mariani *et al*., 2016) and basic text analysis, such as frequency distributions (Xiang *et al*., 2015). After 2017, studies tend to display a wider use of different and more advanced techniques, including more sophisticated sentiment analytics (Alaei *et al*., 2019; Becken *et al*., 2019; Cheng and Jin, 2019; Fu *et al*., 2019; Hao *et al*., 2020; Kirilenko *et al*., 2018; Mehraliyev *et al*., 2020; Park *et al*., 2020; Serrano *et al*., 2020), textual metadata processing (e.g., Miah *et al*., 2017), machine learning techniques (e.g., Ahani *et al*., 2019; Sun *et al*., 2019; Xiang *et al*., 2017; Yin and Wang, 2016), deep learning models (e.g., Chang *et al*., 2020; Ma *et al*., 2018; Zhang *et al*., 2019, 2020), and topic modeling techniques (Mirzaalian and Halpenny, 2019; Moro *et al*., 2019; Sutherland *et al*., 2020). Regardless of the time period considered, only a handful of papers mix a number of different techniques, such as regression analysis, sentiment analysis, machine learning, deep learning and topic modeling (e.g., Xiang *et al*., 2017).

As far as data visualization techniques are concerned, most of the articles deploy traditional tables and figures (e.g., Cheng and Jin, 2019; Kim *et al*., 2019; Ma *et al*., 2018; Mariani and Borghi, 2018; Salas-Olmedo *et al*. (2018).), while a few studies use more advanced visualization techniques, such as density grids (Batista e Silva *et al*., 2018).

5 Discussion and conclusions

5.1 Conclusions

Several key findings emerge from this paper. First, a sharp increase in the number of scientific outputs revolving around BD and analytics can be detected over the last six years (2015–2020). This seems to indicate that the overall amount of knowledge developed on the focal topical area is growing considerably over time, with a rapid acceleration over the most recent years.

Second, and interestingly, the vast majority (72.3% in Scopus and 61.7% in Web of Science) of papers related to BD and analytics in hospitality and tourism settings was published in academic journals whose main reference area is neither hospitality nor tourism (e.g., Chu *et al*., 2019; Preis *et al*., 2020; Renjith *et al*., 2020; Rossetti *et al*., 2014; Sun *et al*, 2016; Toole *et al*., 2015; Wood *et al*., 2013).

Third, when we focus on articles published in H&T outlets, the prevailing macro-topical areas are: 1) the perceptions, experiences, emotions, satisfaction and engagement with hospitality and tourism services of tourists residents and service providers (e.g., Cheng and Jin, 2019; Lee *et al*., 2019; Mariani and Borghi, 2021; Park *et al*., 2020); 2) demand evaluation and forecast/prediction (e.g., Höpken *et al*., 2020; Park and Pan, 2018; Sánchez-Medina and C-Sánchez, 2020); 3) mapping, identification and representation of tourists, tourist behaviors, attractions, destinations and trips (e.g., Batista e Silva *et al*., 2018; Chun *et al*., 2020; Ma *et al*., 2020); 4) knowledge and value creation (Kubo *et al*., 2020; Line *et al*., 2020), also in smart cities and smart ecosystems (Buhalis and Sinarta, 2019); 5) methodological contributions, shedding light on a specific technique or family of techniques (e.g., Alaei *et al*., 2019; Fu *et al*., 2019; Kirilenko *et al*., 2018), measurement problems (e.g., Khalilzadeh and Tasci, 2017), data quality issues (e.g., Xiang *et al*., 2018) or platform issues and features (e.g., Salas-Olmedo *et al*., 2018). These macro-topics have not changed (comparing the period before and after 2017), consistent with the quantitative findings and the cosine similarity index. Interestingly and surprisingly, we did not find a research line on the topic of BD analytics capabilities, which is currently growing fast, especially in the management and information management domain (Gupta and George, 2016; Mikalef *et al*., 2020).

Fourth, most of the H&T outputs are empirical (e.g., Li *et al*., 2020), a few of them explore methodological aspects (e.g., Fu *et al*., 2019), and only a very few of them are conceptual in nature (e.g., Buhalis and Sinarta, 2019). Many of the contributions address a very specific and narrowly defined research question, typically with a limited scope. The conceptual works, however, do not seem to bring BD and analytics under the spotlight, but rather make broader conceptualizations – such as smart destinations (Gretzel *et al*., 2015), smart cities (Batty, 2013), smart services and ecosystems (e.g., Buhalis and Sinarta, 2019) – where BD and analytics are one component cited (often) tangentially, without digging in depth about their practical features and role. Among those studies that have developed conceptual/methodological frameworks, the works address a phenomenon limited in scope, such as information systems at the destination level (Choe and Fesenmaier, 2020); accessibility models (e.g., Järv *et al*., 2018); destination image building processes (e.g., Micera and Crispino, 2017); and strengths and weaknesses of passive mobile data (PMD) (Reif and Schmücker, 2020). There are only two articles that review the BD literature broadly (Li *et al*., 2018; Mariani *et al*., 2018). However, none of these review articles focuses explicitly on the broad and interconnected field of BD and analytics.

Fifth, consistent with the findings of Li *et al*. (2020), the sources of data for the period analyzed (up to 2020) include UGC data (the dominant type, see Ma *et al*., 2020; Park *et al*., 2020;

Salas-Olmedo *et al*., 2018), followed by device data (e.g., Buning and Lulla, 2020; Kubo *et al*., 2020) and last, transaction data (e.g. Gallego and Font, 2020; Park and Pan, 2018). After 2017, there is a growing number of studies mixing data from different sources (e.g., Batista e Silva *et al*., 2018; Park and Pan, 2018; Salas-Olmedo *et al*., 2018).

Sixth, in relation to the size of the data, most of the articles collected less than 1 million records (e.g., Zhao *et al*., 2019). The only two studies that collected more than 70 million records (Gao *et al*., 2013; Zhou *et al*., 2015) are analysis of the H&T context but published in computer science journals. Based on an estimation of the average number of records per dataset used in H&T publications – and therefore, excluding the two aforementioned articles (Gao *et al*., 2013; Zhou *et al*., 2015) – it seems realistic that none of the H&T research team working on BD and analytics would need more than a few dozen terabytes of data to be stored.

Seventh, as far as data collection methods are concerned, most of the researchers using UGC develop their own crawlers, typically using programming languages such as Python, Java, PHP (e.g., Kim *et al*., 2019), with Python slightly becoming the dominant programming language. They also typically leverage API to retrieve data from the major social media platforms such as Facebook, Twitter, and travel review websites like Tripadvisor (e.g., Ma *et al*., 2018). Researchers using device data typically purchase from, or request data directly to, companies owning the data (such as telecommunication companies) (e.g., Kubo *et al*., 2020). Scholars deploying transaction data either purchase them from the owners, data analytics companies or other types of commercial providers (e.g., Gallego and Font, 2020), or retrieve them using crawlers (Liu *et al*., 2018).

Finally, before 2017 studies published in hospitality and tourism journals used a narrow and limited number of non-advanced techniques, such as regression analyses (e.g., Mariani *et al*., 2016) and basic text analysis, such as word frequency distributions (e.g., Xiang *et al*., 2015). After 2017, studies in hospitality and tourism display a wider use of different and more advanced data science techniques, including more sophisticated sentiment analytics (e.g., Hao *et al*., 2020; Park *et al*., 2020), textual metadata processing (e.g., Miah *et al*., 2017), machine learning techniques (e.g., Ahani *et al*., 2019), deep learning models (e.g., Zhang *et al*., 2019), and topic modeling techniques (e.g., Mirzaalian and Halpenny, 2019). Regardless of the time period considered, only a handful of articles mix a number of different techniques such as regression analysis, textual metadata processing, sentiment analysis, machine learning, and topic modeling (e.g., Xiang *et al*., 2017). Furthermore, most of the H&T articles deploy traditional tables and figures for visualization and reporting (e.g., Cheng and Jin, 2019), with only a very few studies deploying more advanced visualization techniques, such as density grids (Batista e Silva *et al*., 2018).

5.2 Theoretical implications

A number of research implications stem from this systematic literature review. First, an increasing number of scholars are claiming to use BD in their research pertaining to hospitality and tourism. In several cases, this does not seem to reflect the reality. Indeed, BD research projects in industrial and commercial settings typically involve the retrieval and processing of very large amounts of data (in the order of petabytes). On average, an H&T academic article deals with no more than 1 million records (e.g., Cheng and Jin, 2019; Liu *et al*., 2017: Mariani *et al*., 2019), which arguably could easily be stored in a few terabytes and would not require very advanced technologies or computational capabilities. Based on more than 2,300 papers retrieved from both databases (Scopus and WoS), it appears that the way the circumlocution "big data" is used in academic circles in H&T is rather far from the way it is meant in industrial and commercial settings. Consequently, there seems to be a lot of upselling within the H&T scholarly domain – and perhaps also within other social sciences – when authors deploy the circumlocution "big data" or select "big data" among their keywords.

Second, in most of the research analyzed, BD and analytics researchers in hospitality and tourism seem to use BD for mere discovery (Mariani and Borghi, 2018) and sometimes to test hypotheses (e.g., Wang *et al*., 2019), rather than for building theory. Our review illustrates that theoretical development (in relation to recognized theories in the social sciences) is very limited, if not absent. Sometimes, extant conceptual works touch BD not tangentially, but to enrich concepts, such as smart tourism (Gretzel *et al*., 2015), smart hospitality (Buhalis and Leung, 2018), and service co-creation (Buhalis and Sinarta, 2019). In future, researchers explicitly using BD and analytics, should make it clear if, and to what extent, they are using theories and make explicit the theories they want to contribute to, but also the theoretical assumptions that guide their analyses. While some methodological efforts have been made recently, especially in the marketing and information system field to reconcile BD with theory (Jimenez-Marquez *et al*., 2019), an increasing number of hospitality and tourism researchers need to further address the issue of the relationship between data and theory (Berente *et al*., 2018). This emerges clearly from our study, as the body of research produced is rather fragmented – as witnessed by the high number of research questions and research topics. While we managed to identify several macro-topics which allowed us to reduce the variance in the topics emerging from the analysis, many research lines do not display solid conceptual and theoretical interconnections and do not set meaningful research agendas. This result seems to corroborate findings of one of the previous literature reviews on the topic (Mariani *et al*., 2018) and reinforces the idea that conceptual development in the area is limited also, today.

Third, the articles surveyed showcase a limited reflection on the theoretical underpinnings of BD and analytics methodologies. Indeed, often researchers deploy ready on-the-shelf algorithms for data, text and sentiment analysis. However, they very rarely elucidate the assumptions upon which those algorithms are built and make sense of the theories (if any) that lead to developing those algorithms. In line with Mazanec (2020), we found very few studies (e.g., Wang *et al*., 2019) that interpreted their results by means of theories of emotions, such as Plutchik's theory (Plutchik, 1980). We would expect more frequent intellectual efforts aimed at discussing critically, and with focused reviews, specific BD and analytics issues or data science techniques (see Alaei *et al*., 2019).

Fourth, the range of analytical techniques adopted to analyze data (especially UGC textual data) is expanding quickly and the last five years have witnessed a growth in the adoption of techniques borrowed from data science. These include: topic modeling techniques (e.g., Vu *et al*., 2020; Xiang *et al*., 2017), textual metadata processing (e.g., Becken *et al*., 2019; Miah *et al*., 2017), sentiment analysis (Aggarwal and Gour, 2020; Hao *et al*., 2020; Mehralyiev *et al*., 2020; Kirilenko *et al*., 2018), ML techniques (Ahani *et al*., 2019; Chang *et al*., 2020; Höpken *et al*., 2020; Sánchez-Medina and Sánchez, 2020), and deep learning models (Chang *et al*., 2020; Hao *et al*., 2020; Ma *et al*., 2018; Zhang *et al*., 2020; Zhang *et al*., 2019). Several recent studies combine some of the aforementioned techniques (e.g., Aggarwal and Gour, 2020). However, in most of the cases the techniques are implemented by leveraging extant software packages and ready on-the-shelf programming libraries, rather than critically evaluating the effectiveness and quality of those analytics. Future research should build on more recent critical approaches to analytics (e.g., Fu *et al*., 2019) to advance the way we make sense of extant techniques within H&T.

Fifth, while BD analytics have been defined as a holistic process to access, store, analyze, and interpret data conducive to the identification of patterns in the data to create value (Wamba *et al*., 2020), it is not always clear the extent to which analytics are used in a descriptive, explanatory, predictive or prescriptive way. In most cases, analytics seem to be used in a descriptive and explanatory way (e.g., Mariani *et al*., 2016; Wang *et al*., 2019; Xiang *et al*., 2017), rather than a predictive way (Höpken *et al*., 2020; Lee *et al*., 2021). There are only few exceptions which adopt either a design science research approach for prediction (e.g., Miah *et al*., 2017), or adopt search engine indexes to predict tourist arrivals (e.g., Dergiades *et al*., 2018; Goel *et al*., 2018; Gunter and Önder, 2016). Generally, authors themselves seldom label the analytics in their study as descriptive, explanatory, predictive or prescriptive. Given the relatively limited development of predictive analytics (Höpken *et al*., 2020), it is not clear to what extent correlational analysis and hypotheses testing, using BD, might provide predictions rather than interpretation of phenomena. In line with what is suggested by Mazanec (2020), researchers should make clear at the outset if their aim is to describe and explain the present/past or predict future trends – in most cases the boundaries between explanation and prediction are vaguely drawn.

Sixth and last, despite research in the wider management literature emphasizing the key role played by BD analytics capabilities (BDAC) by first conceptualizing BDAC (e.g., George and Gupta, 2016) and later analyzing the relationships between BDAC and organizational innovation and performance (Mikalef *et al*., 2019, 2020), the field of hospitality and tourism lacks a solid theoretical development of how, and to what extent, BDAC are being made sense of and used. In the entire sample of papers pertaining to BD and analytics retrieved from Scopus and WoS, we have not found any study dealing with BDAC. An interesting avenue for research might be to analyze if hospitality and tourism firms developing or outsourcing BDAC are in a better position to see and seize business opportunities and, ultimately, overcome their rivals.

5.3 Practical implications

While the primary aim of a literature review is not to generate implications for practice, this study brings about several critical reflections that might lead to practical implications. First, by keeping the scope of the analysis as wide as possible on the topic of BD and analytics in hospitality and tourism, this systematic literature review suggests that scholars interested in analyzing the H&T context, work in different disciplines and sometimes do not talk to each other. Our analysis shows that many scholars who are not publishing in H&T journals, are contributing to the academic debate in the field of BD and analytics – in line with what Che and Tsai (2016) found – and looking for inter-disciplinary collaborations might help address (in a more holistic way) different real-world practical research and business questions.

Second, increasing competition among hospitality and tourism players is urging them to rely on analytics for better decision-making. This is happening in all the leading hospitality and tourism companies, and in travel intermediaries like Booking.com where they have set up specific roles (such as Head of Analytics, Insights and Data). However, and especially in the wake of COVID-19, not all H&T companies have sufficient resources to spend on data analytics, but they could outsource analytics generation to third companies – like some of their counterparts in other industries are currently doing (e.g., Mariani and Fosso Wamba, 2020). This is also happening in academic research to a certain extent (Kubo *et al*., 2020); researchers are outsourcing data retrieval to data analytics companies.

Third, firms should carry out a need analysis in relation to BD and analytics as needs might differ across organizations. Some organizations, like the Charleston Area Convention and Visitors Bureau examined in several BD studies (Pan and Yang, 2017; Park and Pan, 2018), might have specific needs – such as predicting tourism demand accurately. This is clearly confined to a destination (or a specific organization) and the need can be addressed by an academic research center. Accordingly, smaller H&T firms might gain skilled data scientists by building solid relationships with academic and research centers at a local level. This might allow knowledge transfer to smaller organizations, which would also strengthen their data culture.

Fourth, while specialized companies offering business intelligence tools to the H&T sector (e.g., STR) are investing effectively in analytics to support decision-making at the industry level, more efforts should be made by leading technology companies with competence in analytics (e.g., Alphabet) to liaise with them to help grow analytics for the entire sector. This might empower the industry well beyond the limited contributions that academic research can bring about, as is clear from this literature review.

Last, as the digital transformation of business activities and processes is an unrelenting trend, analytics – from big or small data (Kitchin and Lauriault, 2015) – should be increasingly deployed by leading hospitality, tourism and travel companies to build, improve and innovate their business models, as well as enhance and tailor their products. This will be critical, especially in the wake of the COVID-19 pandemic; the outbreak has made it clear that digital data flows are of paramount importance for H&T firms to adapt to external shocks and capture changes almost in real time (Gallego and Font, 2020; Park *et al*., 2020). This implies that organizations (at the national and local level) should support initiatives to provide critical data for managers to act upon.

5.4 Limitations and future research

This study has some limitations. First, we collected data from the two leading academic databases (Scopus and Web of Science) in line with previous bibliometric research. Further research might consider including outputs from Google Scholar, despite the shortcomings discussed above. Second, future research might also try to embed articles produced in 2021.

Beyond its limitations, this work enabled us to identify a few main themes that can help shape future research leveraging BD and analytics in H&T. First, there is an issue of communication that needs to be resolved. Big data and analytics researchers should openly disclose which theories their study is building on or testing. This was missing in most of the studies and, unless resolved, will not allow a systematic generation of knowledge.

Secondly, a new generation of research teams could use data and analytics to generate theoretical developments for the interpretation of phenomena in H&T well beyond the use of correlational analysis and hypotheses testing. More specifically, theoretical developments could be related to recognized theories in the social sciences, and researchers could make explicit the theories they want to contribute to, together with the theoretical assumptions that guide their analyses. In some cases this happened (Park and Pan, 2018), but it is an exception, not a norm.

Third, while there is a discernible trend towards widespread use of more sophisticated analytical techniques, in most of the cases the algorithms used are deployed as black boxes, without questioning the way they work (Alaei *et al*., 2019). Future researchers should make more sense of the way algorithms are built to better interpret and understand their findings and overcome their limitations.

Last, despite that research in the wider management literature has emphasized the key role played by BDAC (e.g., George and Gupta, 2016; Mikalef *et al*., 2020), the field of H&T lacks a solid theoretical development and test of key theoretical advancements in the field of BDAC. This is a gap that future researchers should cover.

References

Acquisti, A., Taylor, C. and Wagman, L. (2016), "The economics of privacy", *Journal of Economic Literature*, Vol. 54 No. 2, pp.442–492.

Aggarwal, S., and Gour, A. (2020), "Peeking inside the minds of tourists using a novel web analytics approach", *Journal of Hospitality and Tourism Management*, Vol. 45, pp.580–591.

Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L. and Weaven, S. (2019), "Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews", *International Journal of Hospitality Management*, Vol. 80, pp.52–77.

Alaei, A.R., Becken, S. and Stantic, B. (2019), "Sentiment analysis in tourism: capitalizing on big data", *Journal of Travel Research*, Vol. 58 No. 2, pp.175–191.

Ali, F., Park, E.(O)., Kwon, J. and Chae, B.(K). (2019), "30 years of contemporary hospitality management: uncovering the bibliometrics and topical trends", *International Journal of Contemporary Hospitality Management*, Vol. 31 No. 7, pp.2641–2665.

Anandarajan, M., Hill, C. and Nolan, T. (2019), "Text preprocessing". Anandarajan, M., Hill, C. and Nolan, T. (Ed.s). *Practical Text Analytics*, Springer, Cham, pp.45–59.

Baggio, R. (2016), "Improving tourism statistics: merging official records with big data", Fuchs, M., Höpken, W. and L'Hagen (Ed.s). *Big Data and Business Intelligence in the Travel and Tourism Industry*. Mid Sweden University, Östersund, pp.89–92.

Barabási, A.L. (2016), *Network Science,* Cambridge University Press, Cambridge, UK.

Batista e Silva, F., Marín Herrera, M.A., Rosina, K., Barranco, R., Freire, S. and Schiavina, M. (2018), "Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources", *Tourism Management*, Vol. 54, pp.101–115.

Batty, M. (2013), "Big data, smart cities and city planning". *Dialogues in Human Geography*, 3(3), 274–279.

Becken, S., Alaei, A.R., Wang, Y. (2019), "Benefits and pitfalls of using tweets to assess destination sentiment", *Journal of Hospitality and Tourism Technology*, Vol. 11 No. 1, pp.19–34.

Benckendorff, P. and Zehrer, A. (2013), "A network analysis of tourism research", *Annals of Tourism Research*, Vol. 43, pp.121–149.

Berente, N., Seidel, S. and Safadic, H. (2018), "Data-driven computationally intensive theory development", *Information Systems Research*, pp. 1–15, available at: https://doi.org/10.1287/isre.2018.0774 (accessed 30 September 2021).

Biswas, B., Sengupta P. and Chatterjee, D. (2020), "Examining the determinants of the count of customer reviews in peer-to-peer home-sharing platforms using clustering and count regression techniques", *Decision Support Systems*, Vol. 135, 113324.

Borgatti, S. P. (2006), "Identifying sets of key players in a social network". *Computational and Mathematical Organization Theory*, Vol. 12 No. 1, pp.21–34.

Brandt, T., Bendler, J. and Neumann, D. (2017), "Social media analytics and value creation in urban smart tourism ecosystems", *Information & Management*, Vol. 54 pp.703–713.

Bryson, S., Kenwright, D., Cox, M., Ellsworth, D. and Haimes, R. (1999), "Visually exploring gigabyte datasets in real time", *Communications of the ACM,* Vol. 42 No. 8, pp.83–90.

Buhalis, D. and Foerste, M. (2015), "SoCoMo marketing for travel and tourism: empowering co-creation of value", *Journal of Destination Marketing & Management*, Vol. 4 No. 3, pp.151–161.

Buhalis, D. and Leung, R. (2018), "Smart hospitality – Interconnectivity and interoperability towards an ecosystem", *International Journal of Hospitality Management*, Vol. 71, pp.41–50.

Buhalis, D. and Sinarta, Y. (2019), "Real-time co-creation and nowness service: lessons from tourism and hospitality", *Journal of Travel and Tourism Marketing*, Vol. 36 No. 5, pp.563–582.

Buhalis, D. and Leung (2018)

Buning, R.J. and Lulla, V. (2020), "Visitor bikeshare usage: tracking visitor spatiotemporal behavior using big data", *Journal of Sustainable Tourism*, Vol. 29 No. 4, pp.711–731.

Cai, H., Jia, H., Chiu, A.S.F., Hu, X. and Xu, M. (2014), "Siting public electric vehicle charging stations in Beijing using big-data informed travel patterns of the taxi fleet", *Transportation Research Part D: Transport and Environment*, Vol. 33, pp.39–46.

Chang, Y.-C., Ku, C.-H. and Chen, C.-H. (2020), "Using deep learning and visual analytics to explore hotel reviews and responses", *Tourism Management*, Vol. 80, 104129.

Chen, C., Ma, J., Susilo, Y., Liu, Y. and Wang, M. (2016), "The promises of big data and small data for travel behavior (aka human mobility) analysis", *Transportation Research Part C: Emerging Technologies*, Vol. 68, pp.285–299.

Chen, L.F. and Tsai, C.T. (2016), "Data mining framework based on rough set theory to improve location selection decisions: a case study of a restaurant chain", *Tourism Management*, Vol. 53, pp.197–206.

Chen, C., Ma, J., Susilo, Y., Liu, Y. and Wang, M. (2016), "The promises of big data and small data for travel behavior (aka human mobility) analysis", *Transportation Research Part C: Emerging Technologies*, Vol. 68, pp.285–299.

Cheng, M. and Jin, X. (2019), "What do Airbnb users care about? An analysis of online review comments", *International Journal of Hospitality Management*, Vol. 76, pp.58–70.

Chu, K.F., Lam, A.Y. and Li, V.O. (2019), "Deep multi-scale convolutional LSTM network for travel demand and origin-destination predictions", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21 No. 8, pp.3219–3232.

Chua, A., Servillo, L., Marcheggiani, E. and Moere, A.V. (2016), "Mapping Cilento: using geotagged social media data to characterize tourist flows in southern Italy", *Tourism Management*, Vol. 57, pp.295–310.

Chugh, R. and Grandhi, S. (2013), "Why business intelligence? Significance of business intelligence tools and integrating BI governance with corporate governance", *International Journal of E-Entrepreneurship and Innovation*, Vol. 4 No. 2, pp.1–14.

Chun, J., Kim, C.-K., Kim, G.S., Jeong, J. and Lee, W.-K. (2020), "Social big data informs spatially explicit management options for national parks with high tourism pressures", *Tourism Management*, Vol. 81, 104136.

Coker, F. (2014), *Pulse: Understanding the Vital Signs of Your Business*, Ambient Light Publishing, Bellevue, WA.

Cox, M. and Ellsworth, D. (1997), "Managing big data for scientific visualization". ACM SIGGRAPH, Vol. 97, pp.21–38.

da Fontoura Costa, L., Rodrigues, A., Travieso, G. and Villas Boas, P.R. (2007), "Characterization of complex networks: a survey of measurements", *Advances in Physics,* Vol. 56 No. 1, pp.167–242.

Davenport, T.H. (2014), "How strategists use 'big data' to support internal business decisions, discovery and production", *Strategy & Leadership*, Vol. 42 No. 4, pp.45–50.

Davenport, T.H. (2017), "How analytics have changed in the last 10 years", *Harvard Business Review*, online at: https://hbr.org/2017/06/how-analytics-has-changed-in-the-last-10-years-and-how-its-stayed-the-same (last accessed September 2021)

Del Vecchio, P., Mele, G., Ndou, V. and Secundo, G. (2018), "Open innovation and social big data for sustainability: evidence from the tourism industry", *Sustainability*, Vol. 10. No. 9, pp.1–15.

Dergiades, T., Mavragani, E., and Pan, B. (2018), "Google trends and tourists' arrivals: emerging biases and proposed corrections", *Tourism Management,* Vol. 66, pp.108–120.

Deza, M.M. and Deza, E. (2016), *Encyclopedia of Distances* (4th ed.), Springer, Heidelberg.

Dursun, A. and Caber, M. (2016), "Using data mining techniques for profiling profitable hotel customers: an application of RFM analysis", *Tourism Management Perspectives,* Vol. 18, pp.153–160.

Erl, T., Khattak, W. and Buhler, P. (2015), *Big Data Fundamentals – Concepts, Drivers and Techniques*, Prentice Hall , New York.

Fan, W. and Gordon, M.D. (2014), “The power of social media analytics”, Communications of the ACM, Vol 57 No. 6, 74–81.

Ferraris, A., Mazzoleni, A., Devalle, A. and Couturier, J. (2019), “Big data analytics capabilities and knowledge management: impact on firm performance. *Management Decision*, Vol. 57, pp.1923–1936.

Fosso Wamba, S., Gunasekaran, A., Akter, S., Ren, S.J.F., Dubey, R. and Childe, S.J. (2017), “Big data analytics and firm performance: effects of dynamic capabilities”, *Journal of Business Research,* Vol. 70, pp.356–365.

Fosso Wamba, S., Kala Kamdjoug, J.R., Epie Bawack, R. and Keogh, J.G. (2020), “Bitcoin, blockchain and fintech: a systematic review and case studies in the supply chain”, *Production Planning & Control,* Vol. 31, pp.115–142.

Fuchs, M, Hopken, W. and Lexhagen, M. (2014), “Big data analytics for knowledge generation in tourism destinations – a case for Sweden”, *Journal of Destination Marketing & Management*, Vol. 3 No. 4, pp.198–209.

Fu, Y., Hao, J.-X., (Robert) Li, X. and Hsu, C.H.C. (2019), “Predictive accuracy of sentiment analytics for tourism: a metalearning perspective on Chinese travel news”, *Journal of Travel Research*, Vol. 58 No. 4, pp.666–679.

Gallego, I. and Font, X. (2020), “Changes in air passenger demand as a result of the COVID-19 crisis: using big data to inform tourism policy”, *Journal of Sustainable Tourism*, Vol. 29 No. 9, pp.1470–1489.

Gandomi, A. and Haider, M. (2015), “Beyond the hype: big data concepts, methods, and analytics”, *International Journal of Information Management,* Vol. 35 No. 2, pp.137–144.

Gao, S., Liu, y., Wang, Y. and Ma, X. (2013), “Discovering spatial interaction communities from mobile phone data”, *Transactions in GIS*, Vol. 17 No. 3, pp.463–481.

García-Palomares, J.C., Gutiérrez, J. and Mínguez, C. (2015), “Identification of tourist hot spots based on social networks: a comparative analysis of European metropolises using photo-sharing services and GIS”, *Applied Geography*, Vol. 63, p.408–417.

George, G., Haas, M.R. and Pentland, A. (2014), “Big data and management”, *Academy of Management Journal*, Vol. 57 No. 2, pp.321–326.

Gerard, G., Osinga, E.C., Lavie, D. and Scott, B.A. (2016), “Big data and data science methods for management research”, *Academy of Management Journal*, Vol. 59 No. 5, pp.1493–1507.

Goel, R., Garcia, L.M.T., Goodman, A., Johnson, R., Aldred, R., Murugesan, M., Brage, S., Bhalla, K., Woodcock, J. (2018), “Estimating city-level travel patterns using street imagery: a case study of using Google Street View in Britain”, *PLOS ONE*, Vol. 13 No.5, e0196521.

Gretzel, U., Sigala, M., Xiang, Z. and Koo, C. (2015), “Smart tourism: foundations and developments”, *Electronic Markets*, Vol. 25, pp.179–188.

Gruss, R., Kim, E. and Abrahams, A. (2020), “Engaging restaurant customers on Facebook: the power of belongingness appeals on social media”, *Journal of Hospitality and Tourism Research*, Vol. 44 No. 2, pp.201–228.

Gunter, U. and Önder, I. (2016), “Forecasting city arrivals with google analytics”, *Annals of Tourism Research*, Vol. 61, pp.199–212.

Gupta, M. and George, J.F. (2016), Toward the development of a big data analytics capability. *Information & Management*, Vol. 53 No. 8, pp.1049–1064.

Halevi, G., Moed, H. and Bar-Ilan, J. (2017), “Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—review of the literature”, *Journal of informetrics*, Vol. 11 No. 3, pp.823–834.

Hao, J.-X., Fu, Y., Hsu, C., Li, X. and Chen, N. (2020), “Introducing news media sentiment analytics to residents’ attitudes research”, *Journal of Travel Research*, Vol. 59 No. 8, pp.1353–1369.

Hasan, S. and Ukkusuri, S.V. (2014), “Urban activity pattern classification using topic models from online geo-location data”, *Transportation Research Part C: Emerging Technologies*, Vol. 44, pp.363–381.

Hasan, S., Schneider, C.M., Ukkusuri, S.V. and Gonzalez, M.C. (2013), "Spatiotemporal patterns of urban human mobility", *Journal of Statistical Physics*, Vol 151, pp.304–308.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. and Ratti, C. (2014), "Geo-located Twitter as proxy for global mobility patterns", *Cartography and Geographic Information Science,* Vol. 41 No. 3, pp.260–271.

Höpken, W., Eberle, T., Fuchs, M. and Lexhagen, M. (2020), "Improving tourist arrival prediction: a big data and artificial neural network approach", *Journal of Travel Research*, in press.

Järv, O., Tenkanen, H., Salonen, M., Ahas, R. and Toivonen, T. (2018), "Dynamic cities: location-based accessibility modelling as a function of time", *Applied Geography*, Vol. 95, pp.101–110.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L. (2019), "Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey", *Multimedia Tools and Applications*, Vol. 78, pp.15169–15211.

Jimenez-Marquez, J.L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J.L. and Ruiz-Mezcua, B. (2019), "Towards a big data framework for analyzing social media content", *International Journal of Information Management*, Vol. 44, pp.1–12.

Kakatkar, C., Bilgram, V. and Füller, J. (2020). Innovation analytics: leveraging artificial intelligence in the innovation process, *Business Horizons*, Vol. 63 No. 2, pp.171–181.

Khalilzadeh, J. and Tasci, A.D.A. (2017), "Large sample size, significance level, and the effect size: solutions to perils of using big data for academic research", *Tourism Management*, Vol. 62, pp.89–96.

Kim, Y., Kim, C.-K., Lee, D.K., Lee, H.-W. and Andrada, R.I.T. (2019), "Quantifying nature-based tourism in protected areas in developing countries by using social big data", *Tourism Management*, Vol. 72, pp.249–256.

Kim, K., Park, O-J., Yun, S. and Yun, H. (2017), "What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management", *Technological Forecasting and Social Change*, Vol 123, pp.362–369.

Kirilenko, A.P., Stepchenkova, S.O., Kim, H. and Li, X. (2018), "Automated sentiment analysis in tourism: comparison of approaches", *Journal of Travel Research*, Vol. 57 No. 8, pp.1012–1025.

Kitchin, R. and Lauriault, T.P. (2015), "Small data in the era of big data", *GeoJournal,* Vol. 80 No. 4, pp.463–475.

Kubo, T., Uryu, S., Yamano, H., Tsuge, T., Yamakita, T. and Shirayama, Y. (2020), "Mobile phone network data reveal nationwide economic value of coastal tourism under climate change", *Tourism Management*, Vol. 77, pp.104010.

Laney, D. (2001), "3D data management: controlling data volume, velocity and variety", *META Group Research Note*, 6.

Law, R., Sun, S., Fong, D.K.C., Fong, L.H.N. and Fu, H. (2016), "A systematic review of China's outbound tourism research", *International Journal of Contemporary Hospitality Management*, Vol. 28 No. 12, pp.2654–2674.

Lee, M., Kwon, W. and Back, K.-J. (2021), "Artificial intelligence for hospitality big data analytics: developing a prediction model of restaurant review helpfulness for customer decision-making", *International Journal of Contemporary Hospitality Management*, Vol. 33 No. 6, pp.2117–2136.

Lee, M., Lee, S.(A). and Koh, Y. (2019), "Multisensory experience for enhancing hotel guest experience: empirical evidence from big data analytics", *International Journal of Contemporary Hospitality Management*, Vol. 31 No. 11, pp.4313–4337.

Li, H., Hu, M. and Li, G. (2020), "Forecasting tourism demand with multisource big data", *Annals of Tourism Research*, Vol. 83, 102912.

Li, J., Xu, L., Tang, L., Wang, S. and Li, L. (2018), "Big data in tourism research: a literature review", *Tourism Management*, Vol. 68, pp.301–323.

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

Liebowitz, J. (Ed.) (2013), *Big data and business analytics*, CRC Press, Boca Raton, FL.

Line, N.D., Dogru, T., El-Manstrly, D., Buoye, A., Malthouse, E. and Kandampully, J. (2020), “Control, use and ownership of big data: a reciprocal view of customer big data value in the hospitality and tourism industry”, *Tourism Management*, Vol. 80, 104106.

Liu, F., Janssens, D., Wets, G. and Cools, M. (2013), “Annotating mobile phone location data with activity purposes using machine learning algorithms”, *Expert Systems with Applications,* Vol. 40 No. 8, pp.3299–3311.

Liu, Y., Teichert, T., Rossi, M., Li, H. and Hu, F. (2017), “Big data for big insights: investigating language specific drivers of hotel satisfaction with 412,784 user-generated reviews”, *Tourism Management*, Vol. 59, pp.554–563.

Ma, S.D., Kirilenko, A.P. and Stepchenkova, S. (2020), “Special interest tourism is not so special after all: big data evidence from the 2017 great American solar eclipse”, *Tourism Management*, Vol. 77, 104021.

Ma, Y., Xiang, Z., Du, Q. and Fan, W. (2018), “Effects of user-provided photos on hotel review helpfulness: an analytical approach with deep leaning”, *International Journal of Hospitality Management*, Vol. 71, pp.120–131.

Mariani, M.M. and Borghi, M. (2018), “Effects of the Booking.com rating system: bringing hotel class into the picture”, *Tourism Management*, Vol. 66, pp.47–52.

Mariani, M.M and Borghi, M. (2019), “Industry 4.0: a bibliometric review of its managerial intellectual structure and potential evolution in the service industries”, *Technological Forecasting and Social Change*, Vol. 149, 119752.

Mariani, M.M. and Borghi, M. (2021), “Are environmental-related online reviews more helpful? A big data analytics approach”, *International Journal of Contemporary Hospitality Management*, Vol. 33 No. 6, pp.2065–2090.

Mariani, M.M. and Fosso Wamba, S. (2020), “Exploring how consumer goods companies innovate in the digital age: the role of big data analytics companies”, *Journal of Business Research*, Vol. 121, pp.338–352.

Mariani, M.M., Baggio, R., Fuchs, M. and Höpken, W. (2018), “Business intelligence and big data in hospitality and tourism: a systematic literature review”, *International Journal of Contemporary Hospitality Management*, Vol 30 No. 10, pp.3514–3554.

Mariani, M.M., Di Felice, M. and Mura, M. (2016), “Facebook as a destination marketing tool: evidence from Italian regional destination management organizations”, *Tourism Management,* Vol. 54, pp.321–343.

Marine-Roig, E. and Anton Clavé, S. (2015), “Tourism analytics with massive user-generated content: a case study of Barcelona”, *Journal of Destination Marketing & Management*, Vol. 4 No. 3, pp.162–172.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M. and López-Cózar, E.D. (2018), “Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories”, *Journal of Informetrics*, Vol. 12 No. 4, 1160–1177.

Mazanec, J.A. (2020), “Hidden theorizing in big data analytics: with a reference to tourism design research”, *Annals of Tourism Research*, Vol. 83, 102931.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. and Barton, D. (2012), “Big data. The management revolution”, *Harvard Business Review,* Vol. 90 No. 10, pp.61–67.

Mehraliyev, F., Kirilenko, A.P. and Choi, Y. (2020), “From measurement scale to sentiment scale: examining the effect of sensory experiences on online review rating behavior”, *Tourism Management*, Vol. 79, 104096.

Miah, S.J., Vu, H.Q., Gammack, J., McGrath, M. (2017), “A big data analytics method for tourist behaviour analysis”, *Information and Management*, Vol. 54 No. 6, pp.771–785.

Mikalef, P., Boura, M., Lekakos, G. and Krogstie, J. (2019a), “Big data analytics and firm performance: findings from a mixed-method approach”, *Journal of Business Research*, Vol. 98, pp.261–276.

Mikalef, P., Boura, M., Lekakos, G. and Krogstie, J. (2019b), “Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment”, *British Journal of Management*, Vol. 30 No. 2, pp.272–298.

Mikalef, P., Krogstie, J., Pappas, I. O. and Pavlou, P. (2020), "Exploring the relationship between big data analytics capability and competitive performance: the mediating roles of dynamic and operational capabilities", *Information & Management*, Vol. 57 No. 2, 103169.

Mirzaalian, F. and Halpenny, E. (2019), "Social media analytics in hospitality and tourism: a systematic literature review and future trends", *Journal of Hospitality and Tourism Technology*, Vol. 10 No. 4, pp.764–790.

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q. and Vespignani, A. (2013), "The twitter of babel: mapping world languages through microblogging platforms", *PLOS ONE,* Vol. 8 No. 4, doi:10.1371/journal.pone.0061981.

Moro, S., Rita, P., Ramos, P., Esmerado, J. (2019), "Analysing recent augmented and virtual reality developments in tourism", *Journal of Hospitality and Tourism Technology*, Vol. 10 No. 4, pp.571–586.

Nam, D., Lee, J. and Lee, H. (2019), "Business analytics adoption process: an innovation diffusion perspective." *International Journal of Information Management*, Vol. 49, pp.411–423.

Newman, M.E.J. (2004), Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, pp.5200–5205.

Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C. and Gonzalez, M.C. (2015), "Urban magnetism through the lens of geo-tagged photography", *EPJ Data Science*, Vol. 4 No. 5, 10.1140/epjds/s13688-015-0043-3

Pan, B. and Yang, Y. (2017), "Forecasting destination weekly hotel occupancy with big data", *Journal of Travel Research*, Vol. 56 No. 7, pp. 957–970.

Park, E., Kang, J., Choi, D. and Han, J. (2020), "Understanding customers' hotel revisiting behaviour: a sentiment analysis of online feedback reviews", *Current Issues in Tourism*, Vol. 23 No. 5, pp. 605–611.

Park, S., Ok, C. and Chae, B. (2016), "Using Twitter data for cruise tourism marketing research", *Journal of Travel & Tourism Marketing*, Vol. 33 No. 6, pp.885–898.

Park, S.Y. and Pan, B. (2018), "Identifying the next non-stop flying market with a big data approach", *Tourism Management*, Vol. 66, pp.411–421.

Plaza, B. (2011), "Google Analytics for measuring website performance", *Tourism Management*, Vol. 32, No. 3, pp.477–481.

Plutchik, R. (1980), "A general pschoevolutionary theory of emotion", Plutchik, R. and Kellerman, H. (Ed.s.), *Emotion: Theory, Research, and Experience: Vol. 1. Theories of Emotion,* Academic, New York, pp.3 – 34.

Preis, T., Botta, F. and Moat, H.S. (2020), "Sensing global tourism numbers with millions of publicly shared online photographs", *Environment and Planning A: Economy and Space*, Vol. 52 No. 3, pp.471–477.

Puiu, D., Barnaghi, P., Tönjes, R., Kumper, D, Intizar Ali, M., Mileo, A., Parreira J.X. …Fernandes, J. (2016), "CityPulse: large scale data analytics framework for smart cities", *IEEE Access*, Vol. 4, pp.1086–1108.

Ranjbari, M., Shams Esfandabadi, Z. and Scagnelli, S.D. (2020), "A big data approach to map the service quality of short-stay accommodation sharing", *International Journal of Contemporary Hospitality Management*, Vol. 32 No. 8, pp.2575–2592.

Raun, J., Ahas, R. and Tiru, M. (2016), "Measuring tourism destinations using mobile tracking data", *Tourism Management*, Vol. 57, pp.202–212.

Řehůřek, R. and Sojka, P. (2010), "Software framework for topic modelling with large corpora", LREC 2010 Workshop on New Challenges for NLP Frameworks, 19–21 May, Valletta, Malta, pp.4550.

Reif J. and Schmücker D. (2020), "Exploring new ways of visitor tracking using big data sources: opportunities and limits of passive mobile data for tourism", *Journal of Destination Marketing and Management*, Vol. 18, 100481.

Renjith, S., Sreekumar, A. and Jathavedan, M. (2020), "An extensive study on the evolution of context-aware personalized travel recommender systems", *Information Processing & Management*, Vol. 57 No.1, 102078.

Rialti, R., Zollo, L., Ferraris, A. and Alon, I. (2019), "Big data analytics capabilities and performance: evidence from a moderated multi-mediation model", *Technological Forecasting and Social Change*, Vol.149, art. 119781.

Rosetti, M., Stella, F. and Zanker, M. (2016), "Analyzing user reviews in tourism with topic models", *Information Technology & Tourism*, Vol. 16, pp.5–21.
Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P. and Harnisch, M., (2015), "Industry 4.0: the future of productivity and growth in manufacturing industries", Boston Consulting Group, 9 April.
Salas-Olmedo, M.H., Moya-Gómez, B., García-Palomares, J.C. and Gutiérrez, J. (2018), "Tourists' digital footprint in cities: comparing big data sources", *Tourism Management*, Vol. 66, pp.13–25.
Sánchez-Medina, A.J. and C-Sánchez, E. (2020), "Using machine learning and big data for efficient forecasting of hotel booking cancellations", *International Journal of Hospitality Management*, Vol. 89, 102546.
SAS Insights (2017), "Big data. What it is and why it matters", available at: https://www.sas.com/en_gb/insights/big-data/what-is-big-data.html (accessed 5 October 2021).
Schuckert, M., Liu, X. and Law, R. (2015), "Insights into suspicious online ratings: direct evidence from Tripadvisor", *Asia Pacific Journal of Tourism Research*, Vol. 21 No. 3, pp.259–272.
Srinivasan, R. and Swink, M. (2018), "An investigation of visibility and flexibility as complements to supply chain analytics: an organizational information processing theory perspective", *Production and Operations Management*, Vol. 27, pp.1849–1867.
Stylos, N., Zwiegelaar, J. and Buhalis, D. (2021), "Big data empowered agility for dynamic, volatile, and time-sensitive service industries: the case of tourism sector", *International Journal of Contemporary Hospitality Management*, Vol. 33 No. 3, pp. 1015–1036.
Sun, S., Wei, Y., Tsui, K.-L. and Wang, S. (2019), "Forecasting tourist arrivals with machine learning and internet search index", *Tourism Management*, Vol. 70, pp.1–10.
Sun, Y., Song, H., Jara, A.J. and Bie, R. (2016), "Internet of Things and big data analytics for smart and connected communities", *IEEE Access*, Vol. 4, pp.766–773.
Sutherland, I., Sim, Y., Lee, S.K., Byun, J. and Kiatkawsin, K. (2020), "Topic modeling of online accommodation reviews via Latent Dirichlet Allocation", *Sustainability*, Vol. 12 No. 5, 1821.
Tenkanen, H., Di Minin, E., Heikinheimo, V. and Hausmann, A. (2017), "Instagram, Flickr, or Twitter: assessing the usability of social media data for visitor monitoring in protected areas, *Scientific Reports*, Vol. 7, 17615.
The Economist (2017), "The world's most valuable resource is no longer oil, but data", 6 May, available at: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (accessed 5 October 2021).
Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A. and González, M.C. (2015), "The path most traveled: travel demand estimation using big data resources". *Transportation Research Part C: Emerging Technologies*, Vol. 58, pp.162–177.
Traag, V.A., Waltman, L. and van Eck, N.J. (2019), "From Louvain to Leiden: guaranteeing well-connected communities", *Scientific Reports*, Vol. 9, 5233.
Tranfeld, D., Denyer, D. and Smart, P. (2003), "Towards a methodology for developing evidence-informed management knowledge by means of systematic review", *British Journal of Management*, Vol. 14, pp.207–222.
Tussyadiah, L.P. and Zach, F. (2016), "Identifying salient attributes of peer-to-peer accommodation experience", *Journal of Travel & Tourism Marketing*, Vol. 34 No. 5, pp.636–652.
Tu, W., Li, Q., Fang, Z., Shaw, S-L., Zhou, B. and Chang, X. (2016), "Optimizing the locations of electric taxi charging stations: a spatial–temporal demand coverage approach", *Transportation Research Part C: Emerging Technologies*, Vol. 65, pp.172–189.
Van Eck, N.J. and Waltman, L. (2010), "Software survey: VOSviewer, a computer program for bibliometric mapping", *Scientometrics*, Vol. 84 No. 2, pp.523–538.
Vieira, E.S. and Gomes, J.A.N.F. (2009), "A comparison of Scopus and Web of Science for a typical university", *Scientometrics,* Vol. 81 No. 2, pp.587–600.

Vilajosana, I., Llosa, J., Martinez, B. and Domingo Prieto, M. (2013), "Bootstrapping smart cities through a self-sustainable model based on big data flows", *IEEE Communications Magazine*, Vol. 51 No. 6, pp.128–134.

Vu, H.Q., Muskat, B., Li, G. and Law, R. (2020), "Improving the resident–tourist relationship in urban hotspots", *Journal of Sustainable Tourism*, Vol. 29 No. 4, pp.595–615.

Waltman, L. (2016), "A review of the literature on citation impact indicators". *Journal of Informetrics*, Vol. 10, pp.365–391.

Wood, S.A., Guerry, A.D., Silver, J.M. and Lacayo, M. (2013), "Using social media to quantify nature-based tourism and recreation", *Scientific Reports*, Vol.3, doi:10.1038/srep02976.

Xiang, Z., Du, Q., Ma, Y. and Fan, W. (2017), "A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism", *Tourism Management*, Vol. 58, pp. 51–65.

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

Xu, F., Nash, N. and Whitmarsh, L. (2020), "Big data or small data? A methodological review of sustainable tourism", *Journal of Sustainable Tourism*, Vol. 28 No.2, pp.144–163.

Xu, X., Wang, X., Li, Y. and Haghighi, M. (2017), "Business intelligence in online customer textual reviews: understanding consumer perceptions and influential factors", *International Journal of Information Management*, Vol. 31 No. 6, pp.673–683.

Yan, E. and Ding, Y. (2012), "Scholarly network similarities: how bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 7, pp. 1313–1326.

Yang, Y., Pan, B. and Song, H. (2014), "Predicting hotel demand using destination marketing organization's web traffic data", *Journal of Travel Research*, Vol. 53 No. 4, pp.433–447.

Ying, S., Chan, J.H. and 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*, Vol. 32 No. 10, pp. 3249–3269.

Zhang, K., Chen, Y. and Li, C. (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*, Vol. 75, pp.595–608.

Zhang, X., Yang, Y., Zhang, Y. and Zhang, Z. (2020), "Designing tourist experiences amidst air pollution: a spatial analytical approach using social media", *Annals of Tourism Research*, Vol. 84, 102999.

Zhao, P. and Hu, H. (2019), "Geographical patterns of traffic congestion in growing megacities: big data analytics from Beijing", *Cities*, Vol. 92, pp.164–174.

Zhao, Y., Xu, X. and Wang, M. (2019), "Predicting overall customer satisfaction: big data evidence from hotel online textual reviews", *International Journal of Hospitality Management*, Vol. 76, pp.111–121.

Zhong, C., Müller Arisona, S., Huang, X., Batty, M. and Schmitt, G. (2014), "Detecting the dynamics of urban structure through spatial network analysis, *International Journal of Geographical Information Science*, Vol. 28 No. 11, pp.2178–2199.

Zhou, X., Xu, C., Kimmons, B. (2015), "Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform", *Computers, Environment and Urban Systems*, Vol. 54, pp.144–153.

Zupic, I. and Čater, T. (2015), "Bibliometric methods in management and organization". *Organizational Research Methods*, Vol. 18, pp.429–472.