Chapter 6
Mining Big Data for Tourist Hot Spots: Geographical Patterns of Online Footprints

Luis Encalada, Carlos C. Ferreira, Inês Boavida-Portugal and Jorge Rocha

Abstract Understanding the complex, and often unequal, spatiality of tourist demand in urban contexts requires other methodologies, among which the information base available online and in social networks has gained prominence. Innovation supported by Information and Communication Technologies in terms of data access and data exchange has emerged as a complementary supporting tool for the more traditional data collection techniques currently in use, particularly, in urban destinations where there is the need to more (near)real-time monitoring. The capacity to collect and analise massive amounts of data on individual and group behaviour is leading to new data-rich research approaches. This chapter addresses the potential for discovering geographical insights regarding tourists' spatial patterns within a destination, based on the analysis of geotagged data available from two social networks.

Keywords Geotagged photos · Geography · Social networks · Big data
Data mining · Spatial analytics

6.1 Introduction

Information and communications technologies (ICTs) enables to advance with new research questions that facilitate a better understanding of ourselves and the

L. Encalada · C. C. Ferreira · J. Rocha (✉)
Institute of Geography and Spatial Planning, Universidade de Lisboa, Lisbon, Portugal
e-mail: jorge.rocha@campus.ul.pt

L. Encalada
e-mail: luisencalada@campus.ul.pt

C. C. Ferreira
e-mail: carlosferreira@campus.ul.pt

L. Encalada · I. Boavida-Portugal
Department of Spatial Planning and Environment, University of Groningen, Groningen,
The Netherlands
e-mail: i.boavida.portugal@rug.nl

K. Koutsopoulos et al. (eds.), *Geospatial Challenges in the 21st Century*,
Key Challenges in Geography, https://doi.org/10.1007/978-3-030-04750-4_6

surrounding environment (Manovich 2011; Dalbello 2011). The last two decades have brought multiple tags related to the vast quantities of data made available by ICTs, e.g., “big data”, “data avalanche” (Miller 2010), “exaflood” (Swanson 2007).

Geographers have been dealing with some of the issues raised by big data (Barnes 2013), questioning its theory and related practices shifts (Floridi 2012; Boyd and Crawford 2012; Crampton and Krygier 2015). Yet, it still has to be done a substantial and continued effort to understand its geographic relevance, as is in the case of the connection between big data and geography (Graham and Shelton 2013).

Despite big data puts challenges to conventional concepts and practices of “hard” sciences, where Geographic Information Science is included (Goodchild 2013; Gorman 2013), the predominance of big data will undoubtedly lead to a new quantitative turn in geography (Ruppert 2013). This is clearly a new paradigm shift in geography research methodologies: a fourth—data-intensive—paradigm (Nielsen 2011).

Geographic technologies are now integrated into social sciences, and promotes the value of geography to a wider audience. There is a growing list of applications of Geographic Information Systems (GIS) that expose its potential for handling the data deluge. Making sense of big data requires both computationally based analysis methods and the ability to situate the results (Berry 2012). Yet, it brings together the risk of plunging traditional interpretative approaches (Gold 2012). The big data era calls for new capacities of synthesis and synergies between qualitative and quantitative approaches (Sieber et al. 2011).

This paradox alliance between “poets and geeks” (Cohen 2010), can be a unique opportunity for geography, stimulating wider efforts to create a bridge over the qualitative–quantitative crater (Sui and DeLyser 2011) and enabling smart combinations of quantitative and qualitative methodologies (Bodenhamer et al. 2010; Daniels et al. 2011; Dear et al. 2011).

The emergence of critical geography, critical GIS and radical approaches to quantitative geography, fostered the idea that geographers are well prepared to combine quantitative methods with technical practice and critical analysis (Lave et al. 2014). This proved to be not quite true, but currently big data opens, specially through data mining, new possibilities for spatial analysis research (Michel et al. 2011) and can extend the limits of quantitative approaches to a wide array of problems usually addressed qualitatively (Lieberman-Aiden and Michel 2011; Michel et al. 2011).

It is a similar case to the rebirth of social network theory and analysis where due to the growing availability of relational datasets covering human interactions and relationships, researchers managed to implement a new set of theoretical techniques and concepts embracing network analysis (Barabási and Pósfai 2016).

Surveys are an example of this new paradigm shift. This method to collect information can exemplify the crisis of those widely used methods facing some difficulties regarding its utility often caused by the decline of response rates, sampling frames and the narrow ability to record certain variables that are the core or geographical analysis, e.g., accurate geographical location (Burrows and Savage 2014). Gradually, self-reported surveys quantifying human motivations and behaviours are being study and compared with non-traditional data (Struijs et al. 2014; Daas et al. 2015).

Such limitations are still more pronounced while considering that: (i) the majority of social survey data is cross-sectional deprived of a longitudinal temporal facet (Veltri 2017); and (ii) most social datasets are rough clusters of variables due to the restrictions of what can be asked in self-reported approaches.

Big Data is leading to advances on both aspects, shifting from static snapshots to dynamic recounting and from rough aggregations to data with high (spatial and temporal) resolution (González-Bailón 2013; Kitchin 2014).

Understanding social complexity requires the use of a large variety of computational approaches. For instance, the multiscale nature of social clusters comprises a countless diversity of organizational, temporal, and spatial dimensions, occasionally at once. Moreover, computation denotes several computer-based tools, as well as essential concepts and theories, varying from information extraction algorithms to simulation models (Cioffi-Revilla 2014; Alvarez 2016).

Big data and its influence on geographic research has to be interpreted in the context of the computational and algorithmic shift that may progressively influence geography research methods. To understand such shift, a distinction between two modeling approaches has to be addressed (Breiman 2001; Gentle et al. 2012; Tonidandel et al. 2016): (i) The data modeling approach that assumes a stochastic model in which data and parameters follow the assumed model; and (ii) The algorithmic approach that considers the data as complex and unknown, and is focused on finding a function that imitates the mechanism of the data-generation process, reducing the statistical model to a function, and keeping out any assumption about the data, e.g., data distribution assumptions (Breiman 2001; Veltri 2017). Whereas the former evaluates the parameters values from the data and then uses the model for information and/or prediction, in the latter there is a move from data models to algorithms properties. Here, what matters the most is an increased emphasis on processes rather than structures.

Big data introduces the possibility of reframing the epistemology of science, presenting two potential paths underpinned by disparate philosophies, Empiricism, and Data-driven science (Kitchin 2014).

6.1.1 Embracing "Big" Changes Beyond Traditional Methods of Data Collection

The label "big data" points to three features, also known as the 3Vs: (i) volume, regarded as the quantity of captured and stored data; (ii) velocity, intended as the quickness at which data can be collected; and (iii) variety, incorporating both structured (e.g., tables and relations) and unstructured (e.g., text and photographs) data (Kitchin 2014; Tonidandel et al. 2016). A fourth V, Veracity, has been added but, as denoted by Kitchin (2014), the term "big data" goes further and describes a type of analytic approach.

This is precisely the type of data created from immense complex systems simulations, e.g., cities (Miller and Goodchild 2015) but a big percentage of it, is provided by sensors and/or software that collect a wide range of social and environmental patterns and processes (Graham and Shelton 2013; Kitchin 2013). The sources of this spatial and temporal data embrace location-aware tools such as mobile phones, airborne (e.g., unmanned aerial vehicles) and satellite remote sensors. Automated data is also generated as digital traces recorded on social media, among others online platforms (Miller 2010; Sui and Goodchild 2011; Townsend 2013).

There is in big data an enormous potential for innovative statistics (Daas et al. 2015). Geolocation data retrieved from mobile phones records can be used to get virtually instant statistics of tourism and daytime/nighttime population (de Jonge et al. 2012). Simultaneously, social media can serve as the background to produce indicators of human mobility (Hawelka et al. 2014). Big data can also be used to replace or complement historical data sources, e.g., surveys, inquiries and governmental data. For instance, inquiries about road usage may become obsolete if detailed traffic data obtained by sensors on the road come to be available (Struijs and Daas 2013).

Part of big data sources, including social media, are made of empirical data and are not intentionally planned for supporting data analysis, i.e., they do not have a clear structure, a well-defined target population and/or proved quality. In this context, it is problematic to make use of statistical methods based on sampling theory, i.e., traditional methods (Kitchin 2013; Daas and Puts 2014a). Specially, the unstructured facet of several of the big data sources makes exponentially difficult to (data)mine significant statistical information. In numerous of these sources, the data explanation and its relations with social phenomenon's are still a very fuzzy field of analysis (Daas and Puts 2014b; Tonidandel et al. 2016).

In a broader perspective, there is another issue regarding the human and technical capacity required for processing and analyzing big data. Contemporary data researchers are probably better prepared than traditional statisticians are. Perhaps the upmost importance is the necessity for a distinct mind-set because big data points toward a paradigm shift (Kitchin 2014), comprising an increased and improved use of modeling practices (Struijs and Daas 2013; Daas and Puts 2014a).

Before big data, random sampling was the main approach to deal with information burden. This method works well, but has its own fragilities: it only performs well if the sampling is representative. Moreover, the sampling basis, i.e., a procedure for numbering and selecting from populations, may be tricky if numbering is performed imperfectly.

Sample data is also very attach to the objective it was first intended. Since randomness is so important it may be difficult to reanalyze the data with different purposes than those for which it was collected (Mayer-Schonberger and Cukier 2014). By the contrary, several of the new data sources do not rely on samples but in populations. Yet, populations have the problem of being tendentially self-selected rather than sampled. For instance, all people having smartphones, all people who engaged "Flickr" or any other social network, or all vehicles traveling in the City of Lisbon between 17:00 and 21:00 on a specific day. In addition, tweets can be a striking source of information (Tsou et al. 2013; Hawelka et al. 2014) but only a part of them are actually geo-located. Despite the specific characteristics of any of these groups may remain to be clarified; it is possible to generalize them to the populations they were sampled from (Encalada et al. 2017).

Nevertheless, some care should be taken since some information people voluntarily provide could not reflect a "real measurement" about their activities (e.g., digital traces of commuting behaviours). Furthermore, selection biases can also occur in the information people volunteer about their surrounding environment. For example, Open Street Map (OSM) is frequently recognized as a popular Volunteer Geographic Information (VGI) venture. Many places around the world, including those in developed countries, have been mapped through OSM with a noteworthy degree of accurateness. Nonetheless, some places such as tourist locations are mapped faster and/or better than others of less interest to OSM users, such as slums (Haklay 2010).

Of course biases also exist in official maps, because the governments (even the developing nation's ones) frequently do not map unconventional settlements such as slums and/or do not update regularly the existent cartography due to budget restrictions. Yet, the biases in VGI maps are probable more subtle. However, VGI platforms such as OSM facilitate tools for data cleaning and validation, so the users (acting as creators and co-creators) are able to remove the fuzziness as much as is conceivable. Goodchild and Li (2012) discussed the challenges regarding the quality of VGI. They concluded that both, traditional and non-traditional geographic information depend on multiple sources and on people expertise to draw together a cohesive image of the landscape. For instance, surface information may be collected from photogrammetry, terrain measurements, historic sources and crowdsourced data. From this synthesis process, the resulting map might be more truthful than any of the original sources by itself, i.e., the all is more than the sum of the parts.

6.1.2 The Analytic Background of (Big)Data Mining

Defenders of big data suggest that it generates thrilling prospects. Though detractors consider it to be more propaganda than reality (Franks 2012; Savitz 2013). Furthermore, big data analysis can be disapproved as a form of "dust-bowl empiricism" (Ulrich 2015; McAbee et al. 2017). Thus, there is a significant breach in our understanding of both the potential and threats of big data (Tonidandel et al. 2016).

Much of the geographic knowledge is based of formal theories, models, and equations that need to be processed in an informal manner. By the contrary, data mining techniques require explicit representations, e.g., rules and hierarchies, with straight access deprived of processing (Miller 2010).

Geography has a history of a relation between law-seeking (nomothetic) and description-seeking (idiographic) knowledge (Cresswell 2013). Wisely, physical geographers get away from these debates, but the nomothetic-idiographic tension keep on in human geography (Sui and DeLyser 2011; DeLyser and Sui 2012; Cresswell 2013). Possibly without surprise, geography has been censured for invalidated theories, results that cannot be reproduced, and a division amongst practice and science (Landis and Cortina 2015). Putka and Oswald (2015) indicate how geography could benefit by implementing the data algorithmic philosophy, and claim that the actual data modeling philosophy prevents the ability to predict results more accurate, and generates models that do not integrate phenomenon's key drivers, without incorporating uncertainty and complexity in a satisfactory manner.

Big data provides chances to detect genuine relations patterns (Dyche 2012). It is realistic to conceive that big data would allow to clarify some of the residual variance. This incremental legitimacy can arise from improved predictors, e.g., Internet footprints (Youyou et al. 2015).

As denoted by Tonidandel et al. (2016), multiple regression is undoubtedly the most widely used statistical approach. Multiple regression assumes that the model being performed is the most correct. Regrettably, the background theories are hardly ever satisfactorily developed to include the most pertinent variables. Also, many often researchers do not even know what can be the missing variables. Hence, researchers test a limited set of variables, and face the possibility to embrace mislaid variables with implications in the model accuracy, and thus in the conclusions drawn from the data (Antonakis et al. 2010).

In opposition to this traditional methodology, the data analytic approach trusts on multiple models or group of models. While the former focuses on selecting the best model and accept that it properly defines the data-generation process, the later analyses all the possible models to be resultant from the existing set of variables and combines the results through a multiplicity of techniques, e.g., bootstrap aggregation, support vector machines, neural networks (Seni and Elder 2010). The subsequent group of models achieves better results, deriving higher accurate predictions (Markon and Chmielewski 2013; Kaplan and Chen 2014). Big data analytics rooted in machine learning techniques can automatically detect patterns, create predictive models and optimize outcomes, facilitating traditional forms of interpretation and theory building. However, in some cases, the new data analytic techniques may not improve outcomes assessed by using more traditional techniques (Schmidt-Atzert et al. 2011).

Big data is supported by a theory platform and it could not be other way. Rather than compare big data to "dust-bowl empiricism", it leads to what Kitchin (2014) describes as data-driven science. Moreover, as denoted by Miller and Goodchild (2015), a data-driven geography may be emerging.

The notion of data-driven science defends that the generation of hypothesis and theory creation resemble an iterative process where data is used inductively. "Dust-bowl empiricism" stopes after the data mining process whereas the process of data-driven science goes on by coupling the inductive and deductive methods (as an iterative process). Hence, it is possible to name a new category of big data research that handles to the creation of new knowledge (Bakshy et al. 2014). Since the inductive process should not start in a theory-less void, preexisting knowledge guides the analytic engine in order to inform the knowledge discovery process, to originate valuable conclusions instead of detecting any-and-all possible relations (Kitchin 2014).

In spatial analysis, the tendency in the direction of local statistics, e.g., geographically weighted regression (Fotheringham et al. 2002) and (local) indicators of spatial association (Anselin 1995), characterize a concession where the main rules of nomothetic geography can evolve on their own way, across the geographic space. Goodchild (2004) sees GIS as a mix of both the nomothetic and idiographic approaches, retained, respectively, on the software and algorithms, and within the (spatial) databases.

Despite it is possible to go for geographic generalizations, space still matters. Both spatial dependency and heterogeneity generate a local context that shapes the processes that occur at the Earth surface. Geography has this believe for many years now, but this has been strengthened by the recent developments in complex systems theory, i.e., local interactions drive to emergent behaviors that are impossible to understand singularly and independently if they are analyzed from a sole (local or global) perspective. The co-created knowledge derived from the interactions between agents within a certain environment links the local and the global perspectives (Miller and Goodchild 2015).

Briefly, there is not a drastic breakdown with the tradition on geography when researchers move to data-driven geography, especially, in applied research. There is a long-lasting confidence regarding the significance of idiographic knowledge per se and its contribution on creating nomothetic knowledge. Even though this confidence is sometimes weak and questioned, data-driven knowledge discovery offers the chance to advance the relationship amongst idiographic and nomothetic geography. Still, despite the fact that complexity theory supports this idea, it advises at the same time that data-driven knowledge discovery may have intrinsic limitations, i.e. emergent behaviour is unpredictable by definition.

6.2 Big Data from Social Media and Its Potential for Spatial Analysis of Urban Tourism Activities

6.2.1 Tracking Tourists' Itineraries: Non-traditional Data Sources

Most of statistical systems supporting the analysis and understanding of the tourism phenomenon in an urban context are based on the use of three indicators: tourist arrivals; overnights; occupation in accommodation units (Heeley 2011). These indicators allow a generic and dynamic reading of the demand flows associated with city tourism. On the other hand, traditional statistical tools and methods can only measure the participation of tourists in "controlled sites" (e.g., museums, hotels, etc.). Both are, however, very limited when a more in-depth analysis of the phenomenon is sought on an intra-urban scale (Ashworth and Page 2011).

Understanding the complex, and often unequal, spatiality of tourist demand in the urban space requires other methodologies, among which the information base available online and in social networks has gained prominence. This, being increasingly georeferenced, allows a more realistic and informed perception about tourist geography(ies) on urban destinations: places of greater/lesser attractiveness; mobility patterns; etc. Such information reveals an advantageous and complementary option to official data (Goodchild and Li 2012), mainly due to its diversity, quantity, timeliness, and continuity.

Greater access to information—facilitated by new Information and Communication Technologies and a profile of tourists seeking more and more frequent online content—coupled with a growing predisposition to share information in social media, have allowed a greater knowledge of the characteristics and behaviour of tourists (Buhalis and Law 2008; Tussyadiah 2012).

Crowdsourced data, coming from social networks, contributes to the understanding of the fruition/consumption of space within urban destinations. The geotagged photos published by users on "Panoramio" and "Flickr" social networks, during their visit to the city of Lisbon, allow us to present a quantitative and geographic reading of urban tourism spatial production and consumption. Particularly, the data extracted from these sources provides meticulous information, of great value, for the identification of places of concentration, in dense and complex areas.

The emergence of Web 2.0 enabled the use of the Internet as a communication channel, by generating a vast collection of digital platforms, as those identified as social media. Social media is defined as any digital platform where users can participate, create and share content. Kaplan and Haenlein (2010) distinguished the following *media*: blogs, content communities, social networks, websites of recommendations and evaluations (Consumer review websites), instant-messaging sites and photo-sharing, etc. (e.g. Viajecomigo, Tripavisor, Twitter, Facebook, Flickr, Panoramio).

The extensive use of the Internet has increased the influence of the content shared in these platforms, on user behaviour and, more specifically, on the behaviour of tourists (MacKay and Vogt 2012; Tussyadiah 2012). Its impact has been significant in the tourism industry (Leung et al. 2013), denoting a growing tendency for tourists to share their experiences by publishing their recommendations, reviews, photos, or videos about a destination, activity, or service, particularly in social networking sites (Buhalis and Law 2008).

ICTs platforms, sensor networks, and wireless communication systems contributes the integration and data exchange. We are experiencing a new era, where information is produced in part by users. This type of information is referred to as User-generated Content (UGC) or Crowdsourced Data (Kaplan and Haenlein 2010), Volunteer Geographic Information (VGI)—most commonly used in the field of geography—or Community-contributed Data (Goodchild 2007; Andrienko et al. 2009).

In the context of geographic information, online content accessible on these media platforms has become part of the set of data sources available for data gathering, overpassing the condition in which the information was produced and distributed exclusively by the official authorities (Sui et al. 2013).

UGC, as opposed to top-down methodologies, has promoted individuals themselves as information generators with high spatial and temporal resolution, boosting the framework of alternatives to track their location (Sui and Goodchild 2011). Georeferenced information constitutes one of the most important types of UGC. Geospatial technologies enabled social networks with positioning and mapping tools, which have led to a massive volume of georeferenced data.

When tourists use their mobile phones, their credit cards, or through access to social networks, leave behind large amounts of digital traces about their activities within a destination (Buhalis and Amaranggana 2014; Hawelka et al. 2014). These digital traces are often openly available. While traditional methods on geographic data collection were based on technically demanding, accurate, expensive and complicated devices, non-traditional sources offer cost-effective information acquired through everyday devices such as mobile phones (Li et al. 2016).

A valuable feature of the UGC grounded on social media is its continuous availability, almost in real time, which means, in most cases, information can be used to analyse current issues which require continuous observation, allowing to change the analytical meaning of a static approach to a more dynamic monitoring process (Sui and Goodchild 2011; Díaz et al. 2012).

This information can be used as a proxy to find patterns in the spatial distribution of visitors within a destination. For instance, several authors have performed analysis based on data extracted from social networks and other online platforms (e.g., wikipedia, wikitravel, and Foursquare) to identify points of tourist interest in different areas of the world (Tammet et al. 2013).

Besides current developments in storing, processing and analyzing this information, it presents some challenges as the lack of assurance regarding its quality (Li et al. 2016), contrary to what happens with information from official sources, since the latter is collected and documented through well-known established procedures (Goodchild 2013). However, this type of data might play a useful role to drive exploratory analysis of a phenomenon (Goodchild and Li 2012).

All this innovation generated by ICTs in terms of data sources has emerged as an additional and complementary support tool for the more traditional (data) sources. Therefore, non-traditional data should not be regarded as a substitute for official data or data collected through traditional scientific methods, but complementary (Goodchild and Li 2012).

The identification of tourist patterns/behaviours/preferences that express themselves through the digital imprints generated in the tourist destination fill a relevant gap in the knowledge about intra-destination mobility and, generally, on the more informal and less documented fruition of the tourist space.

Although the opportunities provided by online information shared by tourists are plenty and prone to unveil some geographical features in a place or a region (through GIS and spatial analysis), this is a recently open field (Zhou et al. 2015) and still hindered by several constraints (data volume, velocity, variety and reliability, among others) and also by some suspicion as a novel research tool, using new information sources in tourism.

The objectives of this chapter are to highlight consistent patterns of tourism production and consumption, in what can configure different tourist geographies of the city of Lisbon, perceived from the analysis of non-traditional data available in social networks platforms, and at the same time try to understand how this new paradigm of big data and mining techniques will affect the future of geographic analysis.

6.2.2 *Urban Analytics: The City of Lisbon and Lisbon Metropolitan Area*

The city of Lisbon, centrally located in the metropolitan area of Lisbon (LMA). The LMA is composed of five touristic official regions, covering the municipalities of Oeiras, Amadora, Odivelas, Loures, Mafra, Almada, and Lisbon (Fig. 6.1).

In an attempt to demonstrate the novel opportunities of non-traditional data as complementary (data)sources for applied research, we refer to six indicators of tourism activity, based on data made available by one of the three Portuguese Mobile Phone Operators, i.e., NOS®, about foreign users visiting and moving around the

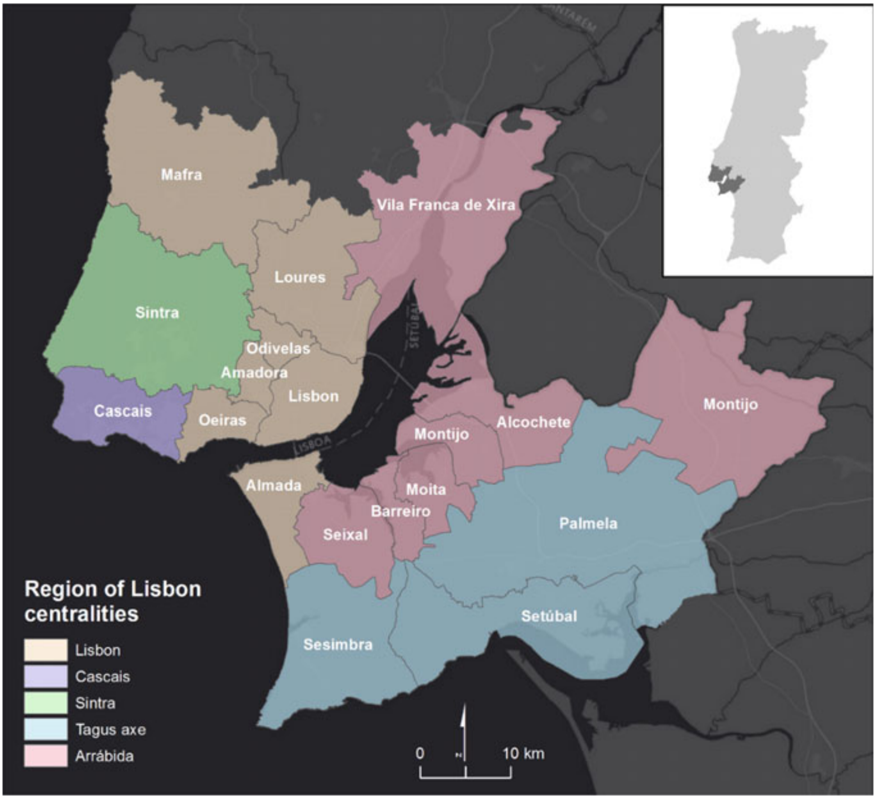

Fig. 6.1 Tourism centralities in the Lisbon metropolitan area

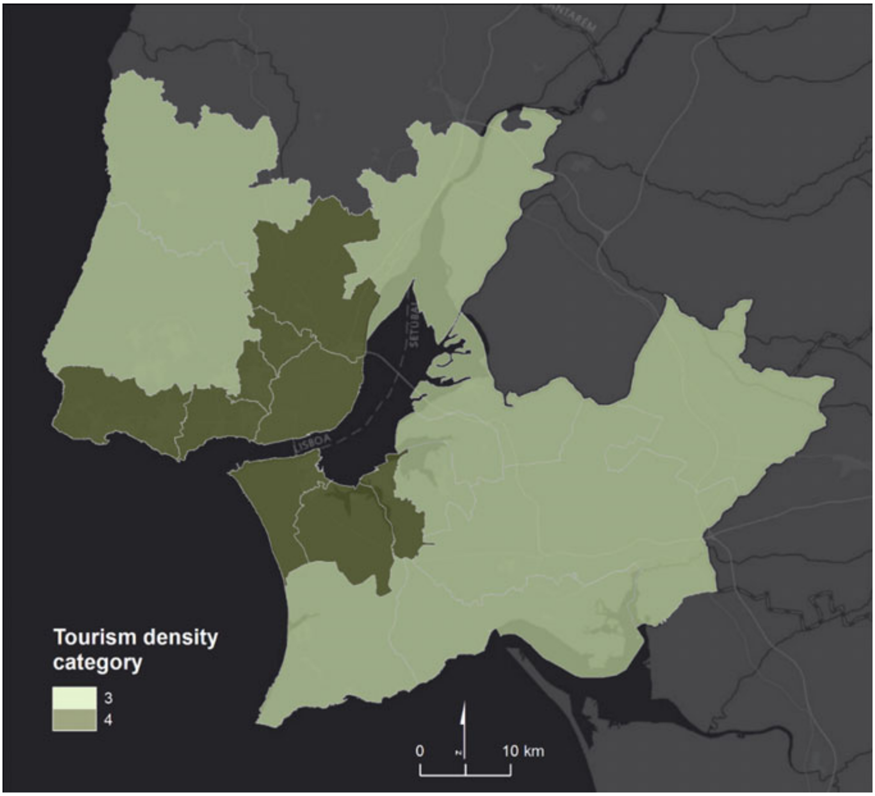

Fig. 6.2 Tourism density by municipality, in Lisbon metropolitan area (2017)

LMA and Portugal. Tourism density values (number of distinct tourists by km^2), an index in a standardized scale of 1 (minimum) to 4 (maximum), show a predominant area within the LMA (Fig. 6.2) which corresponds to the Lisbon centrality. Almost all the cities corresponding to this region (having a tourism density equals to 4) are also in the top ten of Portugal municipalities with highest densities: Lisbon (1); Oeiras (3); Amadora (4); Almada (7); Odivelas (9) and Cascais (10).

As specified by the Tourism Observatory of Lisbon (OTL), "City & short break" is considered the largest motivation for visiting Lisbon (Observatório de Turismo de Lisboa (OTL) 2016). Besides, the weekenders' index (Fig. 6.3) suggest a higher tourist presence on the weekend. This index represents the ratio of the daily average number of tourists at the weekend comparatively to the week, in a month (a value greater than 100 means more tourists on the weekend than on the week). Despite its importance, Lisbon does not take the lead when compared to other cities in the LMA. The two main municipalities are located in the south bank of Tagus River: Sesimbra (156) and Almada (153). They have both a long tradition of sun-and-beach tourism for short periods. This pattern is still supported by the municipalities which

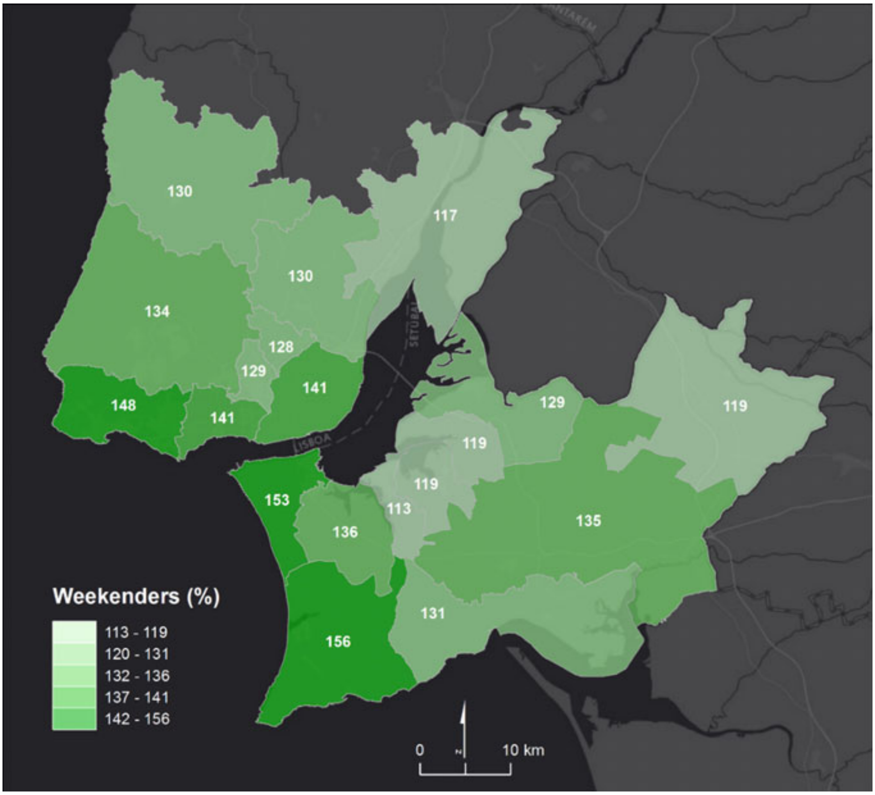

Fig. 6.3 Ratio of weekenders by municipality, in Lisbon metropolitan area (2017)

ranked third and fourth, Oeiras (141) and Cascais (148), both situated in the north Shore of Tagus River and with similar characteristics of the southern counterparts. The only non-beach municipality in the top five is Lisbon (141), which embodies the idea of "City & short break" attractiveness. Yet, these values are very far from the national top ten, denoting the importance of municipalities from the interior, mostly the North interior, of mainland Portugal.

With regards to tourism demand statistics, in 2016, the average stay of foreign guests in Lisbon was up to 2.6 nights. This value is similar for the LMA but lower than mainland Portugal (Instituto Nacional de Estatística (INE) 2017). Besides, the statistic from the mobile phone operator shows a slight increase in 2017, with an average stay of 3.2 nights (Fig. 6.4a). The average number of foreign overnight stays is still one of the lowest in the LMA and from Portugal.

The municipalities leading the national ranking are, as expected, located in the islands of Azores and Madeira. Nonetheless, there are two LMA municipalities in the top ten, i.e., Odivelas (5.1) and Moita (4.6). Both have a vast immigrant population that runs through extended visits from relatives and friends.

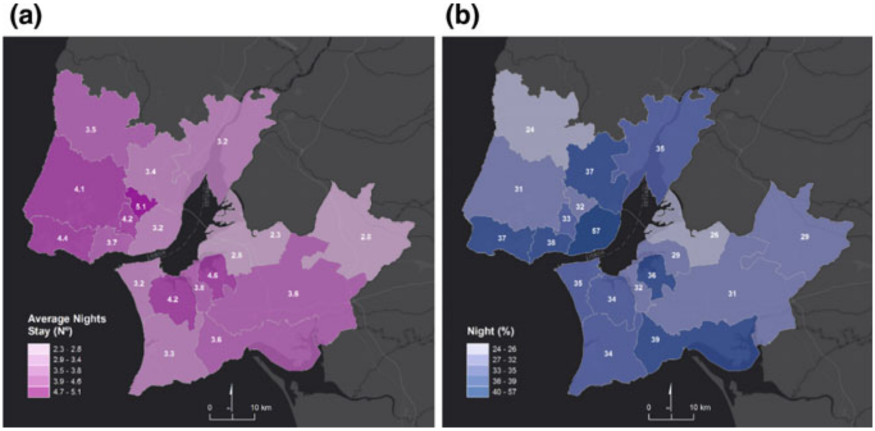

Fig. 6.4 Average number of night's spent (**a**) and night attraction (**b**) by municipality, in Lisbon metropolitan area (2017)

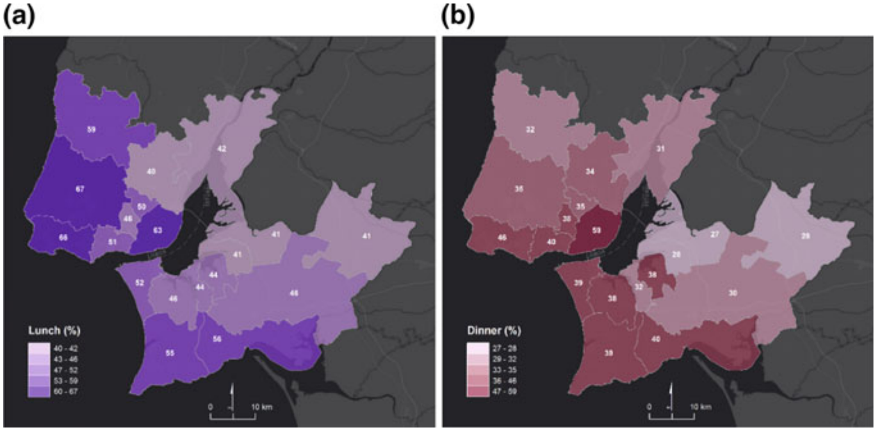

Fig. 6.5 Lunch (**a**) and dinner (**b**) attraction by municipality, in Lisbon metropolitan area (2017)

When considering the night attraction index (it calculates the percentage of tourists at night compared to the total amount of tourists, in a month), Lisbon clearly stands out in the metropolitan area (Fig. 6.4b), being also the second municipality in the national context. Once again, the municipalities with higher night attraction scores are the ones from the Islands. However, in this case, Oporto (tenth) is also in the national top ten.

Finally, lunch (Fig. 6.5a) and dinner (Fig. 6.5b) attraction indexes show different patterns. Both, represent the percentage of tourists at lunchtime/dinnertime in relation to the total number, in a month. Despite having a high lunch attraction (62%), Lisbon stays behind Sintra (67%), a world heritage village, and Cascais (66%), a famous destination for sun-and-beach tourism. On the contrary, when looking at the dinner attraction, Lisbon takes the lead in the metropolitan area. Roughly, values from both indexes may suggest a commuting pattern within the region. It seems that some tourists visiting Lisbon travel to Sintra-Cascais (also known for its landscape-protected area) for spending the day (as denoted by the lunchtime index) but coming back at the end. So far, these outcomes illustrate an interesting flow that should be further explored.

6.2.3 Data Collection of Online Footprints from Social Networks

Recently, researchers have shifted their attention to social media as an alternative data source for collecting information about tourist activities. Here, we explore the value of geotagged data from two social networks in studying tourist spatial behaviour. We use data from "Panoranio" (in 2007 Google acquired Panoramio and closed it down, in late 2016) and "Flickr". Both social networks provide access to the online data through their Application Programming Interfaces (API). In addition to the users' photos (images), metadata information such as users' identification, timestamps and geolocation, is available as well.

According to the protocols from each APIs, the retrieving process must be forwarded through a HTTP request, by setting some parameters (e.g., defining a bounding box to overlap an area, a valid data format, etc.) and data specifications (e.g. a time window, a set of keywords, etc.). Since there are some restrictions to retrieve data (e.g., "Panoramio" allowed to retrieve information up to 500 photos *per* request, for a given area), the requests are usually implemented following an automated scheme (i.e., a recursive algorithm) that controls the iterative process.

The study area was segmented into smaller areas and, for each new unit, we downloaded the online data within its extents. All geotagged photos (and metadata) were stored in a database. "Panoramio" database reached more than 70,000 records (including the image, photo description, users' id, geolocation-coordinates, timestamps, numbers of views, etc.). Similarly, all geotagged photos from "Flickr" were more than 200,000 records.

The identification of photos uploaded by visitors was based on photos' timestamps, following previous works (Girardin et al. 2008; García-Palomares et al. 2015; Encalada et al. 2017). Geotagged photos were classified as belonging to visitors only if the difference (in days) between the timestamps of the first and the last photos uploaded by each user, do not exceed the average stay of foreign tourists within the city. Since the average stay for the last year (2016) of the time-series is close to 3 nights, only photos taken during a period of less than 4 days were cataloged as belonging to visitors.

The final dataset comprise 19,578 photos from "Panoramio" and 73,314 from "Flickr". Photos spatial distribution within the city of Lisbon is depicted on Figs. 6.6 and 6.7 "Panoramio" dataset contains photos from 2008 to 2014, and "Flickr" from

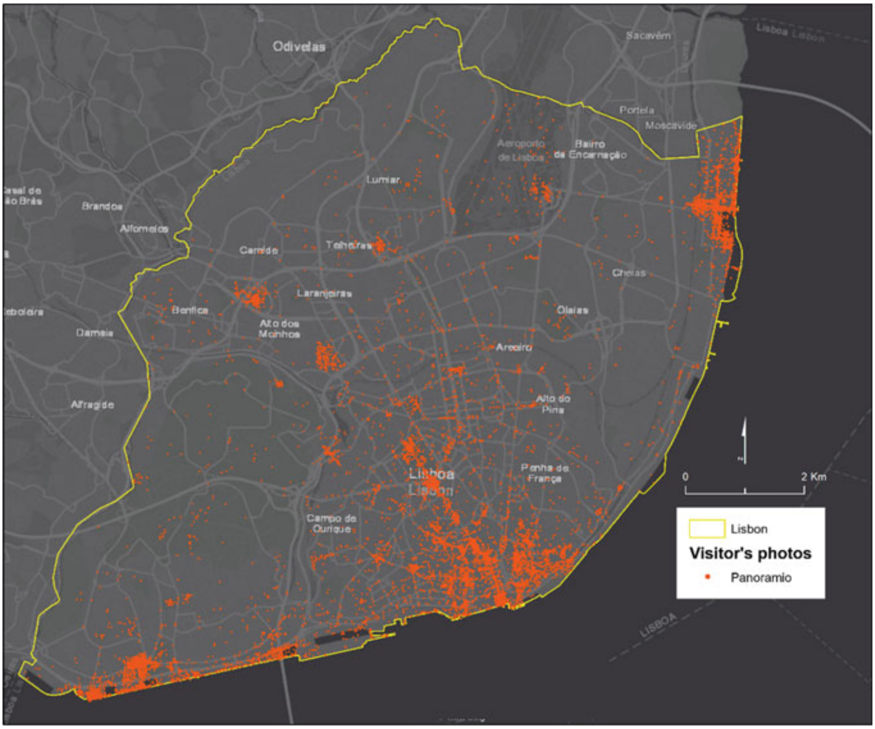

Fig. 6.6 Visitor's geotagged photos from 'Panoramio', from 2008 to 2014

2008 to 2016. These digital footprints belong to more than 15,000 users (from both social networks) considered as city tourist.

6.2.4 Addressing the Spatial Distribution of City Tourists

In this section, we refer to some (traditional) methods of exploratory analysis to discover significant patterns on spatial data. A brief analytical scheme, ranging from global to local metrics, is presented. Our selection criteria regarding spatial analysis techniques relies, mainly, on their generally wider applicability since we aim to reach a broader audience interested on this type of analysis. It should be noted that we keep spatial analysis tools at the simplest level, however, for more details readers should refer to (García-Palomares et al. 2015; Encalada et al. 2017).

When analyzing spatial data, the starting point is to validate the spatial autocorrelation and determine if the global distribution of the data is scattered, concentrated or random. This can be assessed using spatial autocorrelation indexes (e.g. Nearest Neighbor Index, Global Moran's I, Global Getis-Ord Index).

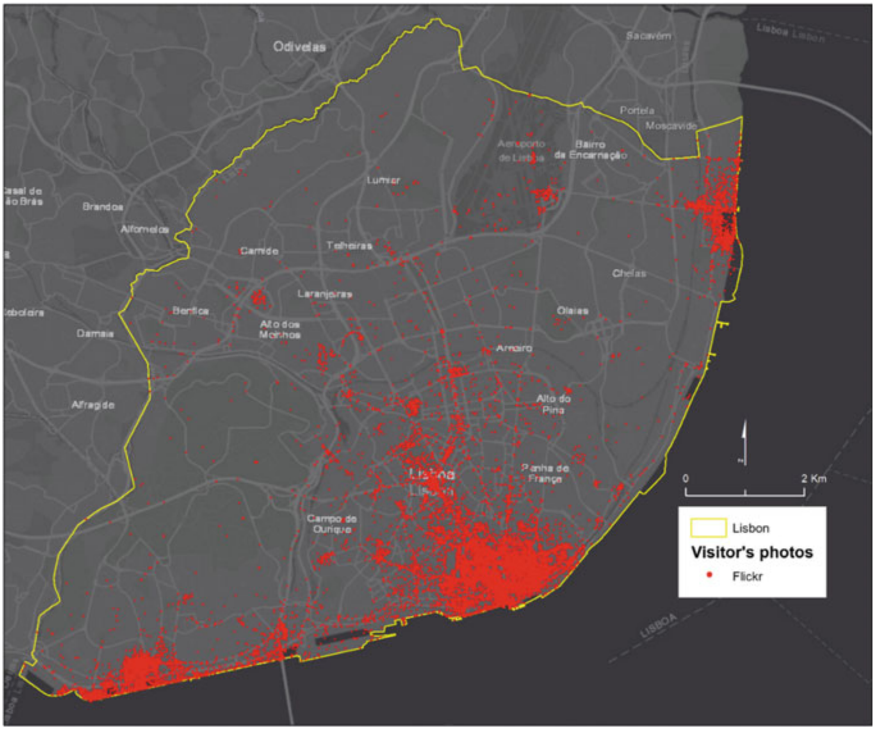

Fig. 6.7 Visitors' geotagged photos from 'Flickr', from 2008 to 2016

Global statistics are not able to measure how spatial dependence varies from place to place. In order to capture the heterogeneity of spatial dependence, statistics might be applied to the local scale. Although the geotagged photos show a roughly concentration on some areas (as depicted on Figs. 5.6 and 5.7), it is necessary to prove whether the observed pattern is statistically significant and, thus, supported by an underlying spatial process. Local indicators such as Local Moran Index and Getis-Ord Gi* lead the disclosure of spatial clusters (i.e. places of tourist concentration) that are statistically significant. The Local Moran Index indicates the spatial concentration of similar values as well as spatial outliers. To perform the analysis, it is necessary to define the neighboring area (i.e., distance threshold) and the nature of the spatial relationship between observations (i.e., the notion of proximity between observations, being reached, in most of the cases, by creating a spatial weights matrix.).

Another practice frequently performed for analyzing and visualizing point features is the Kernel Density Estimation (KDE). It calculates the magnitude per unit area of a given number of points (using the Kernel function), producing a smooth density surface over space by computing the features intensity as density estimation. The basis of these methods is the Tobler’s first law which states that everything is related to everything but near things are more related than distant ones.

Furthermore, spatial patterns change over time. While the former techniques can handle the spatial context, other methods (e.g. Emerging Hot Spot Analysis) support the analysis of both spatial and temporal patterns emerging from a set of observations. Thus, in addition to discover significant spatial clusters, the seasonal pattern can be obtained as well (i.e., whether the hot pots are consecutive, sporadic, etc.).

To assess the geographical patterns of urban tourists within Lisbon, two methods were used, Clusters and Outliers analysis (based on Local Moran Index) and the Kernel Density Estimation. For the Cluster and Outlier analysis, data was aggregated to a continuous hexagon surface. Assuming that all photos within the study area may not be spatially related, a threshold for the neighborhood radius of influence was determined to run the analysis (a threshold distance equals to 150 m). Still, the inverse distance was chosen to conceptualize the spatial relationship, thus, the influence of the neighboring features will decrease as the distance between them increase. Besides, since Kernel function depends on a given distance parameter (e.g., increasing the value of the search radius results in a broader and lower kernel and, thus, showing the spatial tendency on a more global scale), similarly to the Local Moran Index parameters, it was defined a search radius of 150 m. The outcomes are presented in the next section.

6.2.5 Mapping the Spatial Distribution of City Tourists

The visual representation of geotagged photos shows a trend for clustering, mainly in the areas with more touristic appeal. Furthermore, the Local Moran Index outlines a more accurate picture while identifying the city tourist hot spots (i.e., statistically significant places of tourists concentration). Thus, the most relevant touristic sites are discriminated from the overall sample of points previously mapped and depicted on Figs. 5.6 and 5.7.

As expected, clusters are located nearby the well-known city tourist attractions (Figs. 5.8 and 5.9). In general, places of interest such as viewpoints, squares, monumental architecture, and other cultural and recreational attractions function as focus of spatial clusters. The historic center clearly stands out from other touristic areas (on Fig. 5.8, “Eduardo VII” park[3]; “Marquês de Pombal” monument[4]; “Rossio” square[5]; “Comércio” square[6]; “São Jorge” Castle[7]). A smaller number of significant clusters were uncovered over “Belém”—to the southwest (on Fig. 5.8, “Belém” Tower[1]; “Padrão dos Descobrimentos”[2] monument), and in “Parque das Nações”—to the northeast (on Fig. 6.8, Lisbon Oceanarium[8]).

The majority of significant clusters belong to the High-High category, corresponding to places with a large number of tourist’s photos surrounded by similar high counts. On the contrary, there are few atypical clusters (Low-High) located in the surrounding areas of the identified hot spots. These outliers expose some places less visited when compared to the visitor’s presence in its neighborhood.

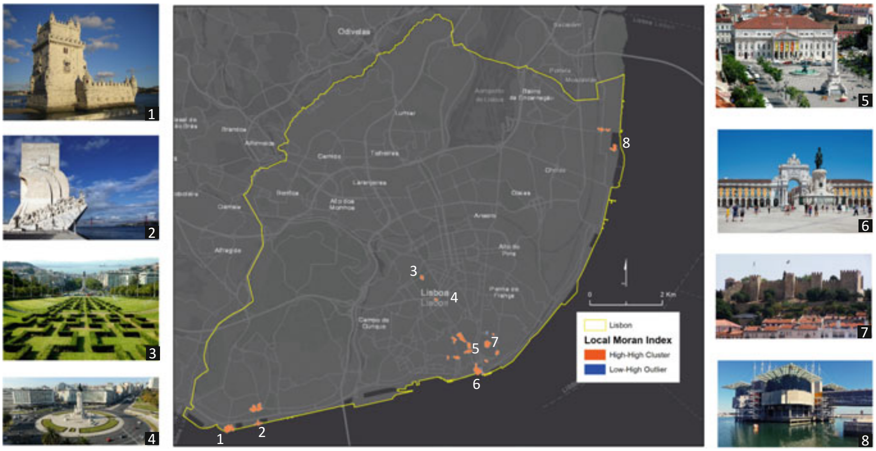

Fig. 6.8 Clusters from 'Panomario' dataset

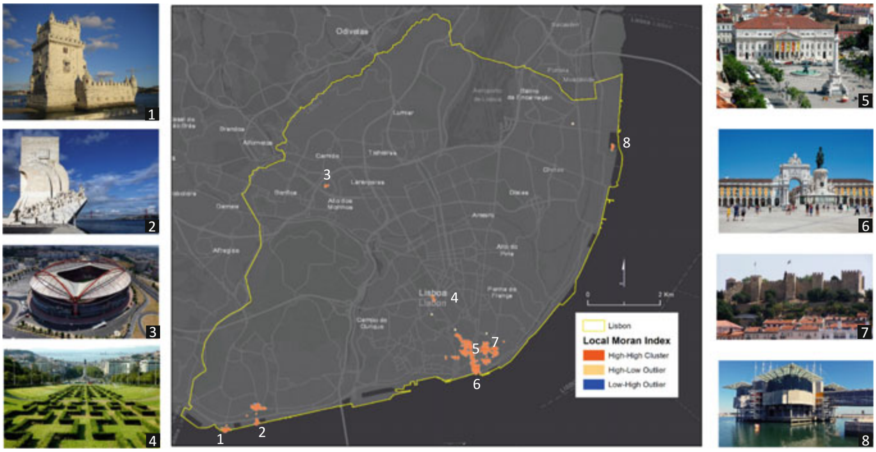

Fig. 6.9 Clusters from 'Flickr' dataset

The main difference between both maps (Figs. 6.8 and 6.9) comes from the fact that "Panoramio" users were more voted to upload photographs illustrating places (e.g., open space areas). Instead, Flickr's photos are more "relaxed" and depict memories about any topic (e.g., social events, daily activities, people, etc.). For instance, the Benfica stadium[3] corresponds to a significant cluster (Fig. 6.9) based on visitors photos from "Flickr", but it is not for "Panoramio".

From the analysis of both density maps (Fig. 6.10), visitor's activity shows a higher density in locals within the Tourism Micro-centralities (i.e., areas of major touristic interest identified by the City Tourism Office). Three areas are highlighted, "Belém" (southwest), the Historic Center and "Parque das Nações" (Northeast). By contrast, few sites with low densities in the inner part of the city reveal an irregular and lower (still significant) attention of city' visitors.

The heat maps effectively summarize some visitor' places of interest. Although "Panoramio" overall kernel density is slightly lower than the results from "Flickr", the hot spots areas match in both cases. For instance, "Comércio" square and "São Jorge" Castle, "Jerónimos" Monastery and "Padrão dos Descobrimentos" monument (southwest), and the Lisbon Oceanarium. Less highlighted places can be identified as well, such as "Rossio" and "Restauradores" squares, and "Marquês de Pombal" monument.

The cross-reading of the resulting maps from the KDE and Cluster analysis, points out that tourist attractions, in fact, are the focus of visitor's clusters. The spatial extent of areas with intensive tourist presence follows a pattern. While the tourist attractions show higher densities, their intensity decrease gradually as the distance to these cores increases. In many cases, the spatial extent follows the physical shape of tourist attractions (e.g., squares, pedestrian streets), expanding across those areas and beyond their perimeter to other nearby areas.

Empirical evidence suggests that urban tourism studies might benefit from geotagged digital data. These outcomes demonstrate that touristic areas can be properly identified and differentiated from others with lesser or non-related tourism activities and visitation.

6.3 Challenges Regarding the New Paradigm Shift in Geography Research

Spatial modeling and in a broad sense Geography, have shifted from a data-scarce to a data-rich environment. The critical change is not about the data volume, but relatively to the variety and the velocity at which georeferenced data can be collected and stored. Data-driven geography is (re)emerging due to a massive georeferenced data flow coming from sensors and people.

Data-driven geography raises some issues that in fact have been long-lasting problems debated within the research community. For instance, dealing with large data volumes, the problem of samples versus populations, the data fuzziness, and the frictions between idiographic and nomothetic approaches. Yet, the conviction that location matters (i.e, spatial context) is intrinsic to geography and serves as a strong motivation to produce refined methods on spatial statistics, time-geography, and GIScience.

Big Data has a huge potential to feed both spatial analysis and modeling, and the geographic knowledge discovery. Nonetheless, there are still some remaining issues, e.g., data validation, non-causal relationship guiding incorrect conclusions, and the creation of understandable data-driven models. The impact of big data and data-driven geography on society remains a current agenda (Mayer-Schonberger and

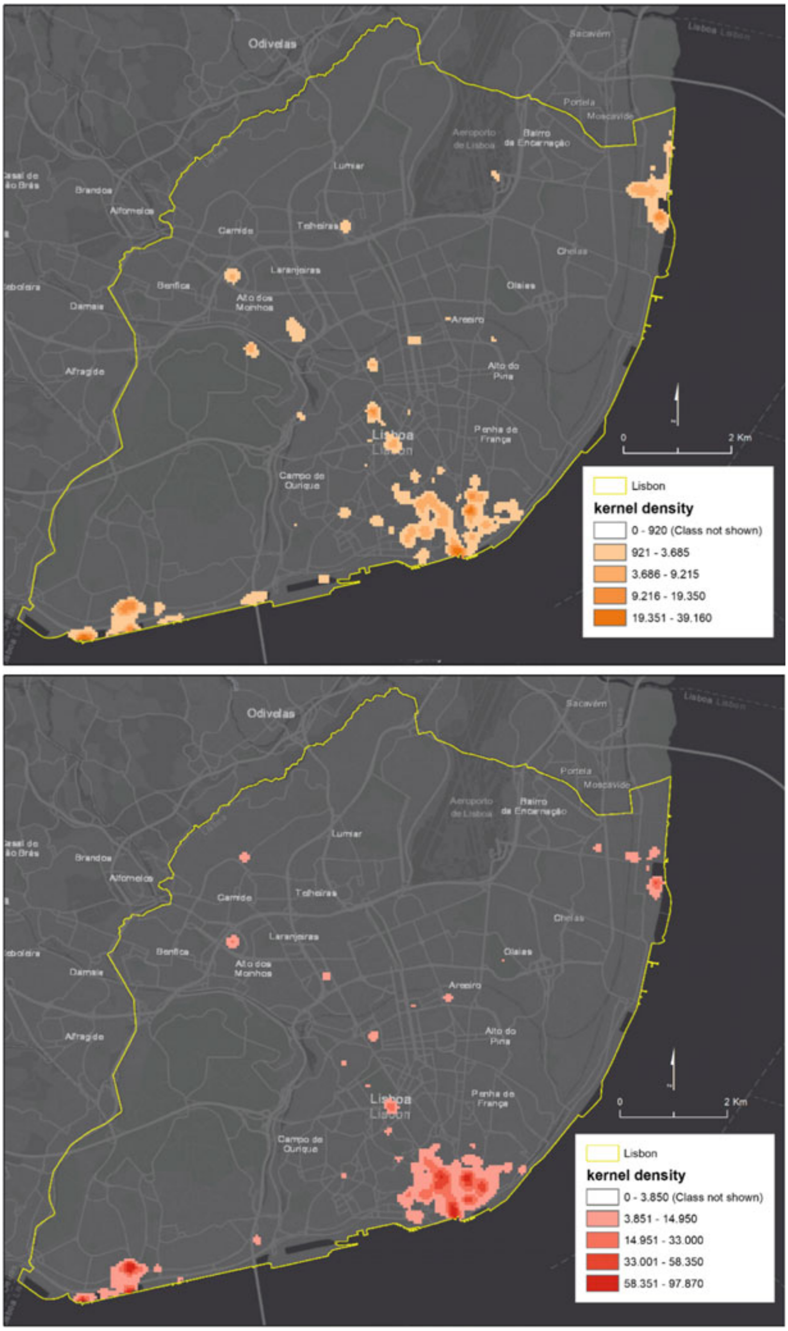

Fig. 6.10 Kernel density (photos/m^2) of geotagged photos from 'Panoramio' (top) and 'Flickr' (bottom)

Cukier 2014). The main concern is with privacy, not only because of the people but also because of the potential repercussions that may stop data-driven research.

Being big data more and more rooted into social-spatial decisions, processes, and institutions, the signifier-signified connection may come to be increasingly fuzzy. As long as we place even more trust in big data, and in the algorithms that are used to produce and analyse it, it is also more likely to lose sight of the big picture of such data represents while distorting the ontological-epistemological boundaries (González-Bailón 2013). This leads to a situation where it is normal to take decisions upon complex data, processed by black-boxed algorithms running in unopen software (Graham 2013).

Still, big data is no danger free, and there is the potential risk of simplifying human agency and the data production frame (Boyd and Crawford 2012; Tinati et al. 2014; Schroeder 2014). Some experiences tell us that, advancements in ICTs, far from being inclusive, often enlarge the socio-spatial roughness of both representation and participation, as evidenced on a variety of online datasets (Graham 2011; Haklay 2013).

Another big data ethical risk derives from what sometimes is designated as machine bias (Angwin et al. 2016). Whereas data are often assumed as objective, big data and the surrounding algorithms may not be. Muñoz and colleagues (2016) show prominent examples of how "bad" data (e.g., badly selected, incomplete, incorrect, or outdated) can lead to discriminatory (biased) outcomes. Keeping that in mind, some attention should be taken, because artificial intelligence can be just as biased as human beings, i.e., discrimination can exist in machine learning.

Indeed, extensive improvements on ICTs have augmented the multimedia narratives about the geographical representation of places, with important implications on the future of geography. Geographers integrating big data with current research paradigms have already transformed and promoted the study of geographical systems and, in the process, have developed new notions of space. This is an opportunity for new research techniques in both the qualitative and quantitative contexts. Big data and data analytics improve the understanding about the consumption of urban space (public or private), leaving physical and digital space, respectively fixed and fluid, while both of them overlap and coexist, each one shaped by the other and its users.

References

Alvarez RM (2016) Computational social science: discovery and prediction. Cambridge University Press, Cambridge

Andrienko G, Andrienko N, Bak P, et al (2009) Analysis of community-contributed space- and time-referenced data (example of flickr and panoramio photos). In: 2009 IEEE symposium on visual analytics science and technology, pp 213–214

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica

Anselin L (1995) Local indicators of spatial association—LISA. Geogr Anal 27:93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Antonakis J, Bendahan S, Jacquart P, Lalive R (2010) On making causal claims:a review and recommendations. Leadersh Q 21:1086–1120. https://doi.org/10.1016/j.leaqua.2010.10.010

Ashworth G, Page SJ (2011) Urban tourism research: recent progress and current paradoxes. Tour Manag 32:1–15. https://doi.org/10.1016/j.tourman.2010.02.002

Bakshy E, Eckles D, Bernstein MS (2014) Designing and deploying online field experiments. In: Proceedings of the 23rd international conference on world wide web. ACM, New York, NY, USA, pp 283–292

Barabási A-L, Pósfai M (2016) Network science, 1st edn. Cambridge University Press, Cambridge

Barnes TJ (2013) Big data, little history. Dialogues Hum Geogr 3:297–302. https://doi.org/10.1177/2043820613514323

Berry DM (ed) (2012) Understanding digital humanities, 1st edn. Palgrave Macmillan UK, Basingstoke

Bodenhamer DJ, Corrigan J, Harris TM (eds) (2010) The spatial humanities. Indiana University Press

Boyd D, Crawford K (2012) Critical questions for big data. Inf Commun Soc 15:662–679. https://doi.org/10.1080/1369118X.2012.678878

Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16:199–231. https://doi.org/10.1214/ss/1009213726

Buhalis D, Amaranggana A (2014) Smart tourism destinations BT—information and communication technologies in tourism 2014. In: Xiang Z, Tussyadiah I (eds) Proceedings of the international conference in Dublin, Ireland, 21–24 January 2014. Springer International Publishing, Cham, pp 553–564

Buhalis D, Law R (2008) Progress in information technology and tourism management: 20 years on and 10 years after the internet—the state of etourism research. Tour Manag 29:609–623. https://doi.org/10.1016/j.tourman.2008.01.005

Burrows R, Savage M (2014) After the crisis? Big data and the methodological challenges of empirical sociology. Big Data Soc 1:2053951714540280. https://doi.org/10.1177/2053951714540280

Cioffi-Revilla C (2014) Introduction to computational social science: principles and applications. Springer-Verlag, London, UK

Cohen P (2010) Humanities 2.0: digital keys for unlocking humanities' riches. New York Times

Crampton J, Krygier J (2015) An introduction to critical cartography. ACME An Int J Crit Geogr 4:11–33

Cresswell T (2013) Geographic thought: a critical introduction. Wiley-Blackwell, Chichester, England

Daas PJH, Puts MJH (2014a) Big data as a source of statistical information. Surv Stat

Daas PJH, Puts MJH (2014b) Social media sentiment and consumer confidence. Frankfurt am Main, Germany

Daas PJ, Marco PJ, Buelens B, van den Hurk PAM (2015) Big data as a source for official statistics. J Off Stat 31:249

Dalbello M (2011) A genealogy of digital humanities. J Doc 67:480–506. https://doi.org/10.1108/00220411111124550

Daniels S, DeLyser D, Entrikin JN, Richardson D (eds) (2011) Envisioning landscapes, making worlds: geography and the humanities. Routledge (Taylor & Francis), Abingdon

de Jonge E, van Pelt M, Roos M (2012) Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. Discussion paper

Dear M, Ketchum J, Luria S, Richardson D (eds) (2011) GeoHumanities: art, history, text at the edge of place. Routledge (Taylor & Francis), Abingdon

DeLyser D, Sui D (2012) Crossing the qualitative-quantitative divide II: inventive approaches to big data, mobile methods, and rhythmanalysis. Prog Hum Geogr 37:293–305. https://doi.org/10.1177/0309132512444063

Díaz L, Granell C, Huerta J, Gould M (2012) Web 2.0 Broker: a standards-based service for spatio-temporal search of crowd-sourced information. Appl Geogr 35:448–459. https://doi.org/10.1016/j.apgeog.2012.09.008

Dyche J (2012) Big data "'Eurekas!'" don't just happen. Harv Bus Rev

Encalada L, Boavida-Portugal I, Cardoso Ferreira C, Rocha J (2017) Identifying tourist places of interest based on digital imprints: towards a sustainable smart city. Sustain 9

Floridi L (2012) Big data and their epistemological challenge. Philos Technol 25:435–437. https://doi.org/10.1007/s13347-012-0093-4

Fotheringham AS, Brunsdon C, Charlton ME (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester, England

Franks B (ed) (2012) Taming the big data tidal wave: finding opportunities in huge data streams with advanced analytics. Wiley, Hoboken, New Jersey

García-Palomares JC, Gutiérrez J, Mínguez C (2015) Identification of tourist hot spots based on social networks: a comparative analysis of European metropolises using photo-sharing services and GIS. Appl Geogr 63:408–417. https://doi.org/10.1016/j.apgeog.2015.08.002

Gentle JE, Härdle WK, Mori Y (eds) (2012) Handbook of computational statistics: concepts and methods. Springer, Berlin, Heidelberg

Girardin F, Fiore FD, Ratti C, Blat J (2008) Leveraging explicitly disclosed location information to understand tourist dynamics: a case study. J Locat Based Serv 2:41–56. https://doi.org/10.1080/17489720802261138

Gold MK (ed) (2012) Debates in the digital humanities. University of Minnesota Press, Minneapolis, MN

González-Bailón S (2013) Big data and the fabric of human geography. Dialogues Hum Geogr 3:292–296. https://doi.org/10.1177/2043820613515379

Goodchild MF (2004) GIScience, geography, form, and process. Ann Assoc Am Geogr 94:709–714. https://doi.org/10.1111/j.1467-8306.2004.00424.x

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69:211–221. https://doi.org/10.1007/s10708-007-9111-y

Goodchild MF (2013) The quality of big (geo)data. Dialogues Hum Geogr 3:280–284. https://doi.org/10.1177/2043820613513392

Goodchild MF, Li L (2012) Assuring the quality of volunteered geographic information. Spat Stat 1:110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Gorman SP (2013) The danger of a big data episteme and the need to evolve geographic information systems. Dialogues Hum Geogr 3:285–291. https://doi.org/10.1177/2043820613513394

Graham M (2011) Time machines and virtual portals: te spatialities of the digital divide. Prog Dev Stud 11:211–227. https://doi.org/10.1177/146499341001100303

Graham M (2013) The virtual dimension. In: Acuto M, Steele W (eds) Global city challenges: debating a concept, improving the practice, 1st edn. Palgrave Macmillan UK, London, pp 117–139

Graham M, Shelton T (2013) Geography and the future of big data, big data and the future of geography. Dialogues Hum Geogr 3:255–261. https://doi.org/10.1177/2043820613513121

Haklay M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey Datasets. Environ Plan B Plan Des 37:682–703. https://doi.org/10.1068/b35097

Haklay M (2013) Neogeography and the delusion of democratisation. Environ Plan A 45:55–69. https://doi.org/10.1068/a45184

Hawelka B, Sitko I, Beinat E et al (2014) Geo-located twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41:260–271. https://doi.org/10.1080/15230406.2014.890072

Heeley J (2011) Inside city tourism: a European perspective. Channel View Publications Ltd, Bristol

Instituto Nacional de Estatística [INE] (2017) Estatísticas do Turismo 2016

Kaplan D, Chen J (2014) Bayesian model averaging for propensity score analysis. Multivar Behav Res 49:505–517. https://doi.org/10.1080/00273171.2014.928492

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of Social Media. Bus Horiz 53:59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kitchin R (2013) Big data and human geography: opportunities, challenges and risks. Dialogues Hum Geogr 3:262–267. https://doi.org/10.1177/2043820613513388

Kitchin R (2014) Big data, new epistemologies and paradigm shifts. Big Data Soc 1:2053951714528481. https://doi.org/10.1177/2053951714528481

Landis RS, Cortina JM (2015) Is ours a hard science (and do we care)? In: Lance CE, Vandenberg RJ (eds) More statistical and methodological myths and urban legends. New York, NY, USA, Routledge (Taylor & Francis), pp 9–35

Lave R, Wilson MW, Barron ES et al (2014) Intervention: critical physical geography. Can Geogr/Le Géographe Can 58:1–10. https://doi.org/10.1111/cag.12061

Leung D, Law R, van Hoof H, Buhalis D (2013) Social media in tourism and hospitality: a literature review. J Travel Tour Mark 30:3–22. https://doi.org/10.1080/10548408.2013.750919

Li S, Dragicevic S, Castro FA et al (2016) Geospatial big data handling theory and methods: a review and research challenges. ISPRS J Photogramm Remote Sens 115:119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Lieberman-Aiden E, Michel J-B (2011) Culturomics: quantitative analysis of culture using millions of digitized books. In: Digital humanities. Stanford University Library

MacKay K, Vogt C (2012) Information technology in everyday and vacation contexts. Ann Tour Res 39:1380–1401. https://doi.org/10.1016/j.annals.2012.02.001

Manovich L (2011) Trending: the promises and the challenges of big social data

Markon KE, Chmielewski M (2013) The effect of response model misspecification and uncertainty on the psychometric properties of estimates. In: Millsap RE, van der Ark LA, Bolt DM, Woods CM (eds) New developments in quantitative psychology: presentations from the 77th annual psychometric society meeting. Springer, New York, New York, NY, pp 85–114

Mayer-Schonberger V, Cukier K (2014) Big data: a revolution that will transform how we live, work, and think, 2nd edn. Eamon Dolan/Mariner Books, London, UK

McAbee ST, Landis RS, Burke MI (2017) Inductive reasoning: the promise of big data. Hum Resour Manag Rev 27:277–290. https://doi.org/10.1016/j.hrmr.2016.08.005

Michel J-B, Shen YK, Aiden AP, et al (2011) Quantitative analysis of culture using millions of digitized books. Science 331(80):176–182

Miller HJ (2010) The data Avalanche is here. shouldn't we be digging? J Reg Sci 50:181–201. https://doi.org/10.1111/j.1467-9787.2009.00641.x

Miller HJ, Goodchild MF (2015) Data-driven geography. GeoJournal 80:449–461. https://doi.org/10.1007/s10708-014-9602-6

Muñoz C, Smith M, Patil D (2016) Big data: a report on algorithmic systems, opportunity, and civil rights. DC, USA, Washington

Nielsen M (2011) Reinventing discovery: the new era of networked science. Princeton University Press, Princeton, New Jersey

Observatório de Turismo de Lisboa (OTL) (2016) Survey to the purpose of trip in Lisbon City 2014

Putka DJ, Oswald FL (2015) Implications of the big data movement for the advancement of IO science and practice. In: Tonidande S, King E, Cortina J (eds) Big data at work: the data science revolution and organizational psychology. New York, NY, USA, Routledge (Taylor & Francis), pp 181–212

Ruppert E (2013) Rethinking empirical social sciences. Dialogues Hum Geogr 3:268–273. https://doi.org/10.1177/2043820613514321

Savitz E (2013) Big data: big hype? Forbes

Schmidt-Atzert L, Krumm S, Lubbe D (2011) Toward stable predictions of apprentices' training success. J Pers Psychol 10:34–42. https://doi.org/10.1027/1866-5888/a000027

Schroeder R (2014) Big Data and the brave new world of social media research. Big Data Soc 1:2053951714563194. https://doi.org/10.1177/2053951714563194

Seni G, Elder JF (2010) Ensemble methods in data mining: improving accuracy through combining predictions. Synth Lect Data Min Knowl Discov 2:1–126. https://doi.org/10.2200/S00240ED1V01Y200912DMK002

Sieber RE, Wellen CC, Jin Y (2011) Spatial cyberinfrastructures, ontologies, and the humanities. Proc Natl Acad Sci U S A 108:5504–5549. https://doi.org/10.1073/pnas.0911052108

Struijs P, Daas P (2013) Big data, big impact? In: Conference of European sattisticians—seminar on statistical data collection—topic (v): integration and management of new data sources. United Nations: Economica Commission for Europe, Geneva, Switzerland, p 9

Struijs P, Braaksma B, Daas PJH (2014) Official statistics and big data. Big Data Soc 1:2053951714538417. https://doi.org/10.1177/2053951714538417

Sui D, DeLyser D (2011) Crossing the qualitative-quantitative chasm I: hybrid geographies, the spatial turn, and volunteered geographic information (VGI). Prog Hum Geogr 36:111–124. https://doi.org/10.1177/0309132510392164

Sui D, Goodchild M (2011) The convergence of GIS and social media: challenges for GIScience. Int J Geogr Inf Sci 25:1737–1748. https://doi.org/10.1080/13658816.2011.604636

Sui D, Goodchild M, Elwood S (2013) Volunteered geographic information, the exaflood, and the growing digital divide. In: Sui D, Elwood S, Goodchild M (eds) Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice. Springer, The Netherlands, Dordrecht, pp 1–12

Swanson B (2007) The coming exaflood. Wall Str J

Tammet T, Luberg A, Järv P (2013) Sightsmap: crowd-sourced popularity of the world places BT—information and communication technologies in tourism 2013. In: Cantoni L, Xiang Z (Phil) (eds) Proceedings of the international conference in Innsbruck, Austria, 22–25 January 2013. Springer, Berlin, Heidelberg, pp 314–325

Tinati R, Halford S, Carr L, Pope C (2014) Big data: methodological challenges and approaches for sociological analysis. Sociology 48:663–681. https://doi.org/10.1177/0038038513511561

Tonidandel S, King EB, Cortina JM (2016) Big data methods: leveraging modern data analytic techniques to build organizational science. Organ Res Methods 1094428116677299. https://doi.org/10.1177/1094428116677299

Townsend AM (2013) Smart cities: big data, civic hackers, and the quest for a New Utopia. W. W. Norton & Company, New York, NY, USA

Tsou M-H, Yang J-A, Lusher D et al (2013) Mapping social activities and concepts with social media (twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. Cartogr Geogr Inf Sci 40:337–348. https://doi.org/10.1080/15230406.2013.799738

Tussyadiah IP (2012) An assessment of contagion on social networking sites BT—information and communication technologies in tourism 2012. In: Fuchs M, Ricci F, Cantoni L (eds) Proceedings of the international conference in Helsingborg, Sweden, 25–27 January 2012. Springer, Vienna, pp 25–35

Ulrich D (2015) Analyzing the analytics agenda, 13 March 2015

Veltri GA (2017) Big data is not only about data: the two cultures of modelling. Big Data Soc 4:1–6. https://doi.org/10.1177/2053951717703997

Youyou W, Kosinski M, Stillwell D (2015) Computer-based personality judgments are more accurate than those made by humans. Proc Natl Acad Sci 112:1036–1040. https://doi.org/10.1073/pnas.1418680112

Zhou X, Xu C, Kimmons B (2015) Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. Comput Environ Urban Syst 54:144–153. https://doi.org/10.1016/j.compenvurbsys.2015.07.006