J. Geogr. Sci. 2020, 30(2): 251-266
DOI: https://doi.org/10.1007/s11442-020-1726-7

Big geodata mining:
Objective, connotations and research issues

PEI Tao[1,2], SONG Ci[1,2], GUO Sihui[1,2], SHU Hua[1,2], LIU Yaxi[1,2], DU Yunyan[1,2], MA Ting[1,2], ZHOU Chenghu[1,2]

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The objective, connotations and research issues of big geodata mining were discussed to address its significance to geographical research in this paper. Big geodata may be categorized into two domains: big earth observation data and big human behavior data. A description of big geodata includes, in addition to the "5Vs" (volume, velocity, value, variety and veracity), a further five features, that is, granularity, scope, density, skewness and precision. Based on this approach, the essence of mining big geodata includes four aspects. First, flow space, where flow replaces points in traditional space, will become the new presentation form for big human behavior data. Second, the objectives for mining big geodata are the spatial patterns and the spatial relationships. Third, the spatiotemporal distributions of big geodata can be viewed as overlays of multiple geographic patterns and the characteristics of the data, namely heterogeneity and homogeneity, may change with scale. Fourth, data mining can be seen as a tool for discovery of geographic patterns and the patterns revealed may be attributed to human-land relationships. The big geodata mining methods may be categorized into two types in view of the mining objective, i.e., classification mining and relationship mining. Future research will be faced by a number of issues, including the aggregation and connection of big geodata, the effective evaluation of the mining results and the challenge for mining to reveal "non-trivial" knowledge.

Keywords: big earth observation data; big human behavior data; geographical spatiotemporal pattern; spatiotemporal heterogeneity; knowledge discovery

1 Introduction

After identifying the challenges and opportunities that massive data would bring to computer science and other disciplines, researchers in computer science some 30 years ago proposed the term data mining. In 1995, Li and Cheng (1995) referred to data mining of spatial data as "knowledge discovery from GIS databases". Much later, Harvey and Han (2009) introduced the term "geographic data mining and knowledge discovery," which marked the substantive intersection of geography and data mining technology. Being an important means for dis-

Received: 2019-08-28 **Accepted:** 2019-09-29
Foundation: National Natural Science Foundation of China, No.41525004, No.41421001
Author: Pei Tao (1972–), Professor, specialized in big geodata mining. E-mail: peit@lreis.ac.cn

covery of geographic patterns, geographic data mining has been widely recognized by geographers. In the subsequent decade, although remarkable progresses have been made in methodological research, little new convincing knowledge regarding geography has been revealed. With the advent of big data, a stream of landmark results have been generated, such as human mobility prediction based on mobile phone data (Song *et al*., 2010), prediction of influenza epidemics using search engines (Ginsberg *et al*., 2009) as well as deep learning algorithms and convolutional neural networks (Silver *et al*., 2016; Silver *et al*., 2017). These developments not only go against common sense, but more importantly, they demonstrate the driving force and momentum brought by application of big data to scientific discovery. Without exception, big data has had a huge impact on geography, forcing geographers to think carefully about key questions such as: What is the nature of big geodata mining? What is the relationship between big geodata and geography? What role can big geodata play in the development of geography? To answer the above questions, this paper considers the concept of big geodata in terms of connotations and denotations, characteristics, core issues and methods. Accordingly, the paper is organized into the following sections. Section 1 explains the connotations and denotations of big geodata. In section 2, the characteristics of big geodata are analyzed in a systematic manner. In the next section, the core issues of big geodata mining are summarized to reveal the nature of the subject area. In section 4, big geodata mining methods are classified in light of the mining tasks. Finally, the development and challenges facing big geodata mining are outlined.

2 Connotations and denotations of big geodata

Although big data has become a hot topic in many disciplines, the definitions regarding the connotations and denotations of big data remain unclear. In fact, the true meaning of defining big geodata is not to explicitly delineate what big geodata is, but to guide how to conduct big geodata analysis and how to overcome the limitations of big geodata in research. Mayer-Schonberger and Cukier (2013) defined the value of big data in their book "Big Data: A Revolution That Will Transform How We Live, Work, and Think". Marr (2015) outlined the "5V" characteristics of big data, that is, volume, velocity, variety, value and veracity. The emergence of big data is mainly due to the rapid advances in sensor technology, networks and computing, thus leading to the key characteristics of large data volume, fast update velocity and wide variety (the first 3Vs). However, big data are uploaded by volunteers on social media platforms (e.g., Facebook, Twitter, Weibo, WeChat) or exist as digital storage records (e.g., mobile phone, bank transactions, utility records). As a result, if using such data as the research object, then big data can be seen as non-purpose observation data. Therefore, such data sources contain much noise, which ultimately leads to low value density and poor veracity (the last 2Vs). In fact, the 5Vs represent only a descriptor of big data, not a definition of big data.

In this paper, the essence of big data is considered to be the "full" coverage of sampling. Note that "full" here does not mean the samples fully cover the object without any space, but that the coverage substantially exceeds that of purposive sampling (here purposive sampling data are called "small data"). The "full" sampling information provided by big data exceeds the limitations of traditional sampling, which inevitably leads to a revolution in the mode of

research. “Full” coverage here involves the dimensions of time, space and attributes. Similar to other fields, big geodata have 5V features, but big geodata also have their own unique features, which will be discussed later. The connotations of big geodata contain at least the following two considerations: first, a distinguishing feature between big geodata and other disciplines lies in whether there are time and space attributes; second, a difference between big geodata and small data is the sample coverage, which was mentioned earlier.

The unique connotations of big geodata originate from the mode of information acquisition, while the denotations are dependent on the means for information acquisition. According to the type of sensors used and the objects recorded, big geodata can be divided into two types: big earth observation data and big human behavior data. Among them, big earth observation data record the characteristics of elements of the earth’s surface, and the sensors are mainly satellite payload (e.g., aerospace or aviation-based) or surface monitoring devices, while the corresponding data include remote sensing imagery, unmanned aircraft imagery and various monitoring (or monitoring network) data. The big earth observation data are acquired mainly in active ways. Big human behavior data record various behavioral activities such as human mobility, socialization and consumption. The sensors are versatile and include mobile handsets, smart cards, social media applications and navigation system-generated signals. Differing from big earth observation data, big human behavior data are acquired mostly in passive ways. The human behavior data can be regarded as footprints of human activity, which include mobile phone signaling data, taxi trajectory data, Internet of Things data and social media data. The focuses of big geo-observation data and big human behavior data are “land” and “human,” respectively. The relationship between human development and geographical environment has always been a central topic in geography, and the advance of big geodata makes it possible to combine these two types of big data, thus providing new resources, new dynamics and new perspectives for the study of human-land relationship in geography. The two types of data focus on different objects, and their data structure, granularity and expression are different, which brings new challenges for big geodata mining.

3 Features of big geodata

The differences between big data and small data have been clarified in the previous section. However, besides the “5V” features, do big geodata have other unique features? The answer to this question could be crucial for analysis. To this end, intrinsic features are discussed from the perspective of the mechanism of generating big geodata. On the one hand, compared to small data, big geodata can be viewed as samples that cover fully the research object. Full coverage includes mainly three aspects: finer granularity, higher density and larger scale. On the other hand, big geodata, especially human behavior big geodata, are typically acquired in a non-purpose way, which may lead to bias and uncertainty. As a result, the features of big geodata can be summarized as spatiotemporal granularity, spatiotemporal scope, spatiotemporal density, spatiotemporal skewness and spatiotemporal precision, which will be explained in the following sections.

3.1 Spatiotemporal granularity

The spatiotemporal granularity is defined as the size of the support unit of geographic in-

formation. The emergence of big geodata makes the granularity smaller. Due to different data acquisition methods, the granularity may vary depending on the different types of data. With regard to big earth observation data, the granularity refers to the pixel size. The minification of the granularity brought by big data can be seen in the increasing refinement of retrieval results for the ground object. For example, enhancement of the resolution of urban images makes the unit of the retrieved information finer, which is changed from the coarse-grained land parcel to the concrete building. In big human behavior data, the granularity refers to the size of the statistical unit (Liu, 2016), and the change of granularity can be seen as the reduction of the statistical unit. Taking demographics as an example, in China's census plan, the census zone is the smallest unit, which is a subdistrict in a city or town in a rural area. The size of the census zone ranges from a few square kilometers to tens of square kilometers, or even larger while the utilization of mobile phone data makes it possible to estimate population in a finer grid. Figure 1 shows the results for a fine scale urban demographic estimated from mobile phone data of Beijing (Liu *et al*., 2018). The unit of the demographic data in Figure 1 is a base station cell (which can be approximated as a Thiessen polygon generated according to the locations of the mobile phone base stations). The scale of the base station cell is about 200 m in the inner city. Similarly, using floating vehicle trajectory data, the assessment of urban road congestion conditions can be refined to any time and any road segment (Zheng *et al*., 2011; Castro *et al*., 2012; Kong *et al*., 2016). When retrieving urban functions using big geodata, the fusion of WeChat location request data, taxi location data, POI data and Quickbird high-resolution images can downscale the functional patches to the building-level (Niu *et al*., 2017); household smart meter information, for example, water consumption, makes it possible to estimate the age, working status and income

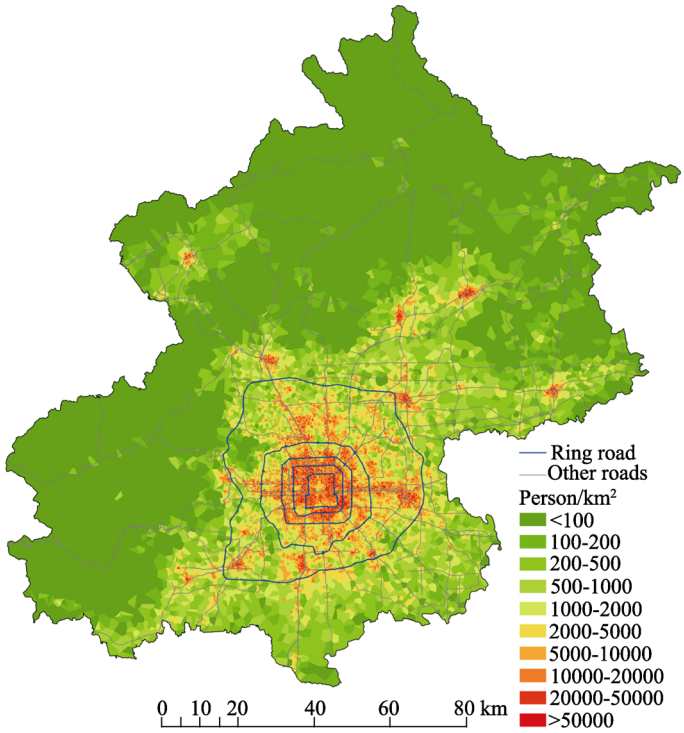

Figure 1 Fine-scale demographic estimation using mobile phone data (Liu *et al*., 2018)

at the family-level (Newing *et al*., 2016). Overall, the refinement of the granularity of big geodata allows us to observe geographical phenomena at a microscale and thus provides new possibilities for studying detailed features and formation mechanisms.

3.2 Spatiotemporal scope

Traditionally, small geodata, due to the high cost of acquisition, are often restricted to local areas, or there is a balance between granularity and scale. In the era of big data, some IT companies take advantage of the Internet to obtain data at large scale meaning at a national or even global level, while at the same time maintaining high resolution. This approach thus ensures that the data cover the large scale and at high resolution. This type of global data product has been applied in many research fields, including global night-time remote sensing data products (NASA, 2017), domestic 30 m resolution global land use data (Chen *et al*., 2015) and global long-term sequence leaf area index products (Liu and Liu, 2015). As to the big human behavior data, the coverage can also be unprecedentedly large, for example, the Spring Festival travel map of China (excluding Taiwan, Hong Kong and Macao) released by Baidu (http://qianxi.baidu.com), the national taxi hot map (excluding Taiwan, Hong Kong and Macao) published by Didi (https://www.didiglobal.com/) and the global user network published by Facebook (http://fbmap.bitaesthetics.com/). From the above analysis, it can be seen that big geodata broadens the scope for research on global change and social laws at the macro level.

3.3 Spatiotemporal density

In addition to the limitation of coverage, traditional geographic research often faces issues regarding the problem of samples of sparse density due to the high cost of sampling. For this reason, the description of geographic distribution based on finite samples usually requires methods of spatial estimation and inference, such as Kriging interpolation (Oliver and Webster, 1990; Stein, 2012), geographically weighted regression (Brunsdon *et al*., 1996; Brunsdon *et al.*, 1998) and environmental factor models (Zhu *et al*., 2010a; 2010b). Although spatial statistical methods can help to reconstruct distributions with finite samples based on spatial autocorrelation, the result cannot replace the true distribution of the attribute. Differing to small data, high-density sampling is one of the fundamental features of big geodata. Specifically, in big earth observation data, with development of sensor technology, the resolution of images has becoming finer and the pixel density has increased, thus resulting in more detailed information being observed. Given the ongoing and continuous upgrading of the global earth observation network, the number of monitoring stations continues to increase. For example, the number of meteorological stations has increased from around 8000 (Zhang, 2010) in the 1960s to more than 100,000 in recent times (NOAA, 2018). The latest figure equates to one observatory per 1490 km^2; the number of Argo buoys for oceanic observation has increased to 3762 as of July 2018 since first deployed in 2000 (Qian and Cheng, 2018). Compared to big human behavior data, the density of traditional “small data” obtained by questionnaires is very low although such sources do have fine granularity. Taking mobile phone data and Tencent location request data as examples, the users effectively cover the majority of the population. With the increasing popularity of smart cards and use of Internet applications, the density of big human behavior data continues to grow; thus with

increase in the density of big data, geographical phenomena will increasingly be observed in more detail.

3.4 Spatiotemporal bias

Although big geodata has demonstrated advantages over "small data" in terms of granularity, scope and density, there are certain flaws, which inevitably incur criticism. Specifically, bias exists ubiquitously in big human behavior data in terms of time, space and attributes. Taking Weibo data as an example, many studies use the data to infer urban functions and human behavior. In fact, serious bias exists in Weibo data mainly with respect to age, gender, space and content. Specifically, the bias due to age is that users of Weibo are mainly those between 18 and 30 years old; and a gender bias is that women are more inclined to use Weibo than men (Yuan *et al*., 2018); a space bias refers to the fact that in China the usage ratio in coastal areas is higher than in the central and western regions. Moreover, Weibo data contain more posts on entertainment, education and finance (Data Center of Sina Micro-blog, 2017) than other topics, such as science and technology. To see how the bias of big geodata affects research, Zhao *et al*. (2016) compared the statistical results of human activities obtained from full samples with that from partially sampled mobile phone data, and found that there was a significant difference between the two sources in terms of the moving distance, the radius of gyration and the mobility entropy. As a result, it is a risk to treat the results derived from biased big data as true facts. Given that the prevalence of bias in big geodata may lead to inappropriate or invalid conclusions, the uncertainties associated with the results should be fully evaluated when using big geodata.

3.5 Spatiotemporal precision

Another aspect that cannot be ignored in big geodata is the low precision. The precision problem is ubiquitous in big data, especially for big geodata, and sometimes can even lead to wrong conclusions being made. Regarding big earth observation data, the precision problem has been extensively studied (Congalton, 1991), and will not be considered here. As to big human behavior data, some of the data are acquired in a passive way (e.g., mobile phone data have been used to estimate the urban population, but have not been collected for that purpose) and some in an active way (e.g., microblog data have been used to measure the mood of the city, but the data are uploaded spontaneously by users). No matter what the mode of data generation is, the data are smeared with various types of errors in terms of space, time and attributes. Taking mobile phone signaling data as an example, due to the limitations of base station capacity, the mobile phone in use may connect to other free base stations rather than the nearest base station. Thus, a spatial error is generated if we locate the handset in the cell of the nearest base station. Similarly, in social media data, the location, time and content of events uploaded by users are not necessarily a true description of the event. Therefore, unlike purposive sampling data, the errors in big geodata do not only originate from technical issues, but also from uncontrollable factors, or even deliberate factors (Zhao and Zhang, 2018). The existence of errors in big geodata often leads to incomplete or even wrong knowledge being generated. Google's success and failure in flu prediction is one such notable example (Ginsberg *et al*., 2009; Lazer *et al*., 2014).

The impact of big geodata comes from its fine granularity, large scope and high density,

and this is difficult to obtain from traditional small data, while the shortcomings of bias and low precision can be compensated for by use of small data. Therefore, big geodata and small data have their own advantages and disadvantages and one cannot completely replace the other. The combination of the two may draw on the strong points of big data to offset any weaknesses. In the application of big geodata, more attention should be paid to its limitations so as to avoid the generation of errors and abuse of the data.

4 Core issues of big geodata mining

As we know, the value of data lies in the knowledge hidden in it (Benz *et al*., 2004; Fayyad *et al*., 1996). To discover knowledge from data, data mining is essential. How can data mining techniques be used to reveal the geographical knowledge hidden in big geodata? To answer this question, we need to clarify four aspects. The first aspect concerns the expression of big geodata. Big earth observation data are conventionally structured such that the data are obtained from well-designed sensor systems. Differing from that, big human behavior data are obtained via stored records of data which originate from different human behaviors, and there are different types and forms of behavior which are difficult to be structured. Therefore, data expression is a prerequisite in big geodata mining. The second aspect is the definition of big geodata mining which may clarify the objective and essence of mining. As already mentioned, big geodata are complex in terms of expression, structure and content, which means that only when the objective and essence of mining are ascertained, can big geodata mining be developed into a branch of geographical information science or even an independent discipline. The third aspect concerns the scalability of big geodata. Having features such as fine granularity, wide scope and high density, big geodata contain multi-scale geographical information incomparable with traditional small data. In such cases, the question of how to deal with the scalability of big geodata mining should be clarified before processing big geodata. Last, but not least, the relationship between big geodata mining and geography needs to be clarified. Faced with big geodata, a new information source for geographic research, it is important to determine the relationship between big geodata mining and geographical science, especially the role that big geodata mining plays in the practice of research.

4.1 Expression of big geodata: location space and flow space

The objects that big earth observation data focus are on the surface of the earth, while the subjects of big human behavior data are human. The interaction between the earth's surface and human being can be viewed as relationships between the subjects and the environment. The basis of earth observation data is location, on which the variation of geographical attributes can be observed and measured. In geographical research, earth observation data can be viewed as containing the environmental factors which affect human behavior and activities. This location-based data can be expressed in the framework of location space, where location is the basic element and Euclidean distance is the basic measurement (Sun and Lu, 2005). In location space, location is the basic unit for the expression of geographical features and geographical phenomena appear as the instantaneous state of geographical features (Han *et al*., 2011). Location space is the expression framework of traditional maps and locations and the spatial relationships are the essence of the spatiotemporal pattern. One major objec-

tive of big geodata mining is to reveal the patterns in location space.

Big human behavior data are a reflection of human activities. In human-related activities and interactions, flow can be seen as a basic unit (flow can be defined as a point pair containing an origin (O) point and a destination (D) point) which includes human flow, commodity flow, information flow, capital flow and relationship flow. A flow can be thought of as an interaction between two nodes (locations) (Castells, 1999; Goodchild *et al*., 2007). For big human behavior data, the distance between O and D is no longer the only measure of their relationship, with this coexisting with various other measures such as time, cost and attractiveness (Batty, 2013). In human behavior data, the focus on humans is not limited to changes in locations but includes various travel behaviors and social relationships. The weakening of the distance effect and the complex flow patterns make the location space unable to adapt to the expression and analysis of big human behavior data. To solve this problem, a new conceptual space regarding flow is needed. Here, we define it as flow space, where the flow is seen as a basic unit, and multiple flows between different nodes forms a network. In the flow space, the core is the interaction between locations, and the purpose of data mining in flow space is to extract the interaction patterns. However, at present, it is difficult to express complex flow patterns on a traditional map. To achieve this, a holographic map and virtual reality technology will become a new analytical framework for describing flow patterns. The flow space is different from location space in terms of measurement, nature and the analytical model, therefore, research on the extraction method for the spatiotemporal flow pattern is an important direction of big geodata mining.

4.2 Content of big geodata mining: patterns and relationships

The purpose of geographical data mining is to find rules and exceptions between geographical objects as well as between geographical objects and the environment. Accordingly, the content of big geodata mining can also be divided into two parts: the first part is the mining of geographical spatiotemporal patterns, the essence of which is to discover the spatiotemporal distributions of geographical objects; the second is the mining of geographical spatiotemporal relations, the essence of which is to discover the relationships between the geographical objects and their environmental factors. Due to the distinct features of big geodata, the contents of mining big geodata are different from those of "small data".

4.2.1 Spatiotemporal patterns in geography

The currently acknowledged theorems in geography are the first law of spatial correlation and the second law of spatial heterogeneity (Tobler, 1970; Goodchild, 2004). The meanings of the two theorems seem to be opposite, but actually these two theorems jointly describe geographical phenomena from two standpoints: objects that are near to each other are considered similar, but they are still different from each other. In location space, the first law describes the relationship between object similarity and distance, while heterogeneity describes spatial non-stationarity. In flow space, spatial correlation is represented by the existence of the spatial network structure, that is, the flow of close origins and destinations constitutes the connections between different locations, and the strength of the connection is related to different variables such as distance. The heterogeneity is represented by the discrepancy of flows between the spatial locations. The essence of spatiotemporal pattern mining of big geodata is to reveal the "similarity and heterogeneity" rules and the resulting spa-

tiotemporal distribution caused by spatiotemporal correlation and heterogeneity. The so-called "heterogeneity" refers to the difference between geographical objects, and the "similarity" refers to the commonality of different objects. Taking the pattern mining of seismic data as an example, on the one hand, it is necessary to determine the "heterogeneity-similarity" rules, based on which clustered and background earthquakes can be distinguished; on the other hand, on the basis of the "heterogeneity-similarity" rule found, the spatial distribution and characteristics of clustered earthquakes and background can be determined. The former is related to the inference of the "heterogeneity-similarity" rule, specifically, "similar" earthquakes may belong to the same statistical distribution, and the "heterogeneous" earthquakes may be divided into different statistical distributions; the latter is related to the extraction of spatiotemporal distribution, but in fact, the spatiotemporal distributions of clustered earthquakes and background can be seen as a comprehensive result of the law of spatial correlation and that of heterogeneity. The main tasks of traditional geographic data mining include the discrimination of spatiotemporal heterogeneity, the extraction of geographic abnormal patterns, the identification of spatial distribution patterns and the extraction of geographic evolution trends, while the changes brought by big geodata are concentrated on the types and scales of the pattern. As to the type, in addition to heterogeneity and distribution of grid, point and field, big geodata mining will place emphasis on complex patterns such as structure and heterogeneity of spatiotemporal sequence, flow and network; as to the scale, given the characteristics of fine granularity, wide scope and high density, big geodata mining will result in more macroscopic, more comprehensive and even finer patterns.

4.2.2 Geographical spatiotemporal relationship

The relationships between geographic objects and environmental factors usually appear as correlations or associations. Correlation is often used to characterize the quantitative relationship between attributes of geographic objects and their environmental factors, say the relationship between the level of soil lead (Pb) pollution and its proximity to highways (Du *et al*., 2007); while association often describes the dependencies between geographic objects that exist or occur simultaneously, such as the relationship between theft and burglary cases (Chen *et al*., 2015). Two factors should be clarified when addressing the geographical spatiotemporal relationship. Taking leaded gasoline and the relationship between lead pollution and the proximity to highway as an example of the interaction between variables, here the exhaust emissions of vehicles on highways can lead to an increase in the lead content in the soil adjacent to the highway; another example is the relationship between the change in soil lead content and its distance to the source of pollution, that is, the closer the soil is to the highway, the higher the lead content. The changes brought by big geodata mainly lies in the change of relationship types and the transformation between different relationships. Specifically, on the one hand, the types of relationships between variables are more diverse and complex compared to those of small data, while nonlinear, uncertain, and multivariate spatiotemporal relationships become one of the core issues of big geodata mining (Cheng *et al*., 2018); on the other hand, except for spatiotemporal relationship mining under the same space, the inversion of information between different spaces (such as social space, physical space, and emotional space) has become one of the main features of big data mining. Such relationship transformations between different spaces have also become the core of big data

thought; for instance, the inversion of economic conditions from remote sensing data (Keola *et al*., 2015), the utilization of "hot" search words to predict influenza trends (Ginsberg *et al*., 2009) and the inference of urban land use from mobile phone data (Pei *et al*., 2014).

It should be noted that compared with small data, big geodata may demonstrate "stronger" spatiotemporal correlation. As a result, it is relatively easy to "discover" various spatiotemporal relationships from them. Given that the causes of these relationships are often very complicated, whether there is a causal relationship or not needs to be carefully checked for. Taking the co-occurrence of theft and burglary cases as an example, the actual reason for their co-occurrence may be that the natural and social environment in a particular region is poor, thus resulting in high incidences of various types of crimes, however, there is no obvious causal relationship between these various crimes (Chen *et al*., 2015).

4.3 Inner structure of the geographic pattern: scale and superposition

As mentioned above, the purpose of big geodata mining is to extract the spatiotemporal patterns and the spatiotemporal relationships. Numerous studies have shown that geographical patterns, distributions and processes are all scale dependent. In other words, any geographic pattern can occur at a certain scale, so big geodata mining is inseparable from scale. Specifically, the purpose of pattern mining is to identify the groups such that objects are similar for the same group and different between different groups. The seeming contradiction between heterogeneity and homogeneity can be transformed by a change of scale. For instance, Figure 2 displays the transformation from heterogeneity to homogeneity at different observation scales for point process data; that is, a point pattern can be seen as heterogeneous at a large scale (Figure 2a), while at a small scale a local part of the pattern can be viewed as homogeneous (Figure 2b). Thus, large-scale complex patterns can be viewed as the superposition of several local homogeneous patterns. Similarly, the spatiotemporal relationship in big geodata is also scale dependent, in other words, the scale of the geographic elements determines the scale of correlation in them. Specifically, the variation at a large scale determines the overall trend while the variation at a small scale determines the local correlation. The superposition of patterns at different scales finally forms the overall complex relationship. For example, the topographic and climatic factors at a large scale determine the macroscopic pattern of China's population distribution, while the characteristics of mesoscale factors (e.g., local landscape, topography, traffic) determine the local distribution of the population. Multi-scale superposition of multiple factors ultimately leads to a complex spatial distribution of geographic phenomena.

The scalability in data mining can be seen as the difference between patterns mined at different observation scales. Based on this, Pei *et al*. (2012; 2013) proposed the decomposition theory of point processes. The main idea behind the theory is that at a given point geographical phenomena can be regarded as the superposition of n homogenous point processes of different densities. Specifically, on the one hand, the spatiotemporal distribution can be seen as the overlay of homogenous processes at different scales; on the other hand, the occurrence of the point process phenomenon can also be viewed as the outcome of the superposition of factors at different scales. For example, the seismicity in a region can be viewed as the superposition of background earthquakes, fault induced earthquakes and local seismic sequences. At the same time, the occurrence of earthquakes can also be treated as the out-

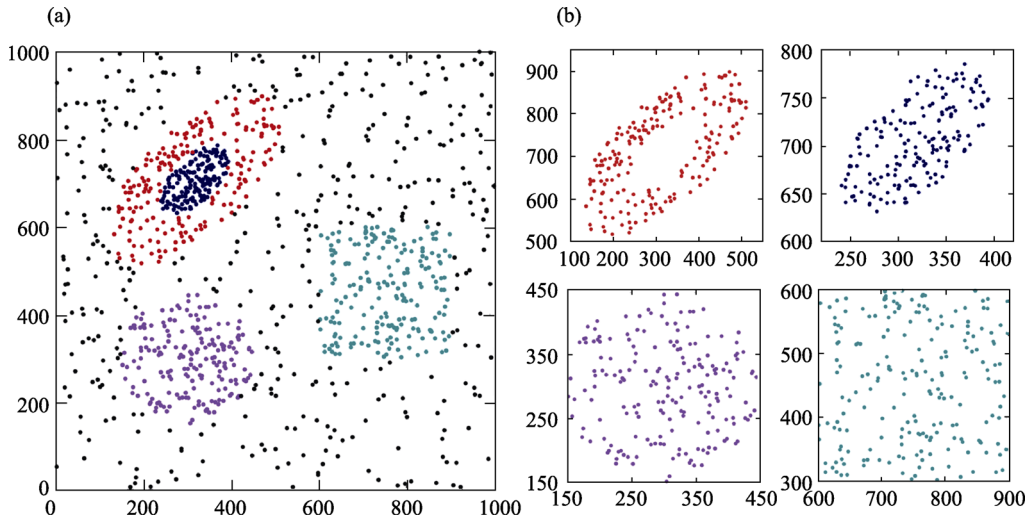

Figure 2 Transformation between homogeneity and heterogeneity of the geographical point process at difference scales: (a) heterogeneity at a large scale; (b) homogeneity at a small scale.

come of the superposition of tectonic movements at different scales. The above analysis suggests that usually multi-scale features exist in big geodata ranging from fine to coarse features. Given that geographical patterns and their mechanisms are thought of as the outcome of comprehensive superposition, then, in turn, big geodata mining can be seen as the process of decomposing patterns and relationships.

4.4 Knowledge discovered from big geodata: human-environment relationships behind geographical patterns

Geodata mining can generally be divided into two stages, the discovery of non-trivial patterns and the exploration of the causes of the patterns. Concerning big geodata mining, the purpose of big earth observation data mining is to uncover the patterns of elements of the earth's surface, while that of big human behavior data mining is to reveal the human behavior patterns. After that, a key question is what are the underlying mechanisms behind the patterns? Big geodata, especially the emergence of big human behavior data, constitutes the complete condition for revealing the mechanism of geographic patterns from the perspective of human-environment relationships. Landuse patterns, reflecting the characteristics of elements of the earth's surface, are definitely the result of human behavior. Mobile phone call data record communication behavior and the data may vary for different areas, but the data indeed contain information about the urban function (Pei *et al*., 2014). The patterns behind big geodata can be attributed to the relationships between humans and the environment. Specifically, the environmental patterns imply human factors while the patterns in human behavior are affected by the environment. Therefore, deriving the human-environment relationships in geographical patterns is a connotation of big geodata mining.

Geographic research has evolved from the first paradigm to the fourth paradigm, i.e., the empirical paradigm (first paradigm), the positivistic paradigm (second paradigm), the system simulation paradigm (the third paradigm) and the data-driven paradigm (the fourth paradigm) (Cheng *et al*., 2018). Current geographic research is more dependent on big geo-

data and the associated analytical methods. Big geodata mining has been recognized as an important tool for the discovery of geographical knowledge. Note that although big geodata mining can generate "knowledge" such as geographic patterns and their relation to the environment, the question of whether there is causality in the relationships as well as authenticity in the patterns of "knowledge" does require strict verification through simulation, observation and experimentation.

5 Big geodata mining methods

Given the dramatic growth of big geodata, various data mining methods have been developed to discover the non-trivial knowledge. These methods can be categorized in different ways according to whether there is a dependency or not on prior knowledge, that is, the methods can be categorized as a model-driven approach or a data-driven approach (Niemeijer, 2002; Miller and Goodchild, 2015). According to the mining tasks, methods can be categorized into spatial clustering, spatial classification, spatial association rules mining, spatial serial rules mining, spatial dependency rules mining, spatial outlier detection, spatial hotspot and trend analysis (Li *et al*., 2001; 2002; Mennis and Guo, 2009; Shekhar *et al*., 2011). According to the mining objects, they can be classified into relational data mining, object-oriented data mining, image data mining, text data mining, multimedia data mining and network data mining (Chen *et al*., 1996; Džeroski, 2009). According to the mining models, they can be grouped into machine learning methods, statistical methods, neural networks and database methods (Jun Lee and Siau, 2001; Wang *et al*., 2005; Reddy, 2011). In this paper, the methods are classified according to the mining objective. The first category is classification mining, which includes spatial clustering (Ester *et al*., 1996; Han *et al*., 2009), spatial classification (Koperski *et al*., 1998), spatial decision tree (Friedl and Brodley, 1997) and point process decomposition (Pei *et al*., 2012). The classification mining methods are generally used to differentiate geographic objects and then to extract the spatial-temporal patterns. The second is relationship mining, such as association rules mining (Koperski and Han, 1995; Huang *et al*., 2004), principal component analysis (Byrne *et al*., 1980; Novembre *et al*., 2008) and regression analysis (Beale *et al*., 2010; McMillen, 2004). These approaches are usually used to determine the relationship between different spatial-temporal variables, especially the relationship between geographic objects and the environment. Moreover, there are some methods that can be assigned into either classification mining or relationship mining, such as neural networks (Atkinson and Tatnall, 1997; Li and Yeh, 2002), support vector machine (Pal and Mather, 2005; Brereton and Lloyd, 2010) and random forest (Gislason *et al*., 2006; Mutanga *et al*., 2012). Besides data mining methods, some optimization methods are also used extensively to estimate the coefficients of the mining model, such as the expectation-maximization algorithm (Moon, 1996) and the Markov chain Monte Carlo (MCMC) algorithm (Andrieu *et al*., 2003). Meanwhile, given the complexity of geographical phenomena, recently artificial intelligence (AI) methods, such as deep learning (LeCun *et al*., 2015), have been used extensively in big geodata mining. However, the AI methods do not belong to big geodata mining methods and only when combined with other data mining methods, can they be seen as part of the data mining toolkit.

Due to the "5V" and other five features of big geodata, several problems need to be ad-

dressed when applying traditional data mining methods to big geodata mining. First, large volumes of data pose a serious challenge to data mining methods. Big data equate to high computational cost, therefore, how to conduct parallel and distributed computing is an essential problem that needs to be solved. Second, the emergence of big data often produces two types of effects when applying the methods to complex questions. On the one hand, some models can be simplified as the volume of data increases. For example, traditional methods using the shortest path analysis mainly depend on the complex optimization model, while, with the emergence of big GPS data, the recoding of the trajectories of floating cars makes it a simple query (Yuan *et al*., 2013; Dai *et al*., 2015). On the other hand, big geodata will result in the development of more complex models. For example, the emergence of big GPS data generates problems like "how to share taxis in order to save energy" (Vazifeh *et al*., 2018). Third, due to the problems of bias and precision, evaluation needs to be performed rigorously to ensure the validity of the results.

6 Conclusion

The coming of the big data era has significantly influenced the development of geographic research. As a special type of big data, the challenges that big geodata mining faces mainly originate from three sources. The first is the aggregation of multi-source big geodata. Big geodata are varied in terms of granularity, presentation and structure. How to realize "vertical" fusion through joining non-spatial attributes and "horizontal" integration by extending the spatiotemporal scope will become the key to mining big geodata in the future. The second is uncertainties caused by the bias and precision of big geodata. How to evaluate and apply the mining results is an unavoidable challenge. Third, producing "non-trivial" knowledge is a tough task for big geodata mining. Current research has generated some notable achievements in statistical physics and AI, nevertheless, the role, which data mining plays in the geographic sciences, has not been widely recognized as yet. For instance, although the uneven pattern of demographics in China can be revealed by examination of the Tencent location request data (https://heat.qq.com/index.php), and as spectacular as it is, this basic pattern was revealed as the "Hu Line" several decades ago.

Facing the challenges mentioned above, the future of big geodata mining seems not difficult to predict. First, big geodata mining should address and solve the basic problems of geographic science at the large scale with a finer granularity and with more comprehensive data. Global change and its impact on society, human behavior and its relationship to the environment, the social characteristics of the earth's surface and urban dynamics are considered the hotspots for future study. Second, big geodata mining methods will evolve by adapting themselves to the "5Vs" and the other five features mentioned above. On the one hand, only more efficient and robust algorithms can be adapted to undertake the complex big geodata mining tasks; on the other hand, new AI methods, including training with large volume samples, will introduce new expectations for solving complex geographical problems. Third, research on big earth observation data will extend from the traditional observations of the earth's surface to the perceptions of social activities, and may bring more scientific and commercial applications, whereas research on big human behavior data will extend from the perception of social activities to the retrieval of earth surface features, and become

more widely used in urban studies. The combination of these two types of big data will eventually form a breakthrough to reveal the human-land relationships in geographic sciences.

References

Andrieu C, De Freitas N, Doucet A *et al*., 2003. An introduction to MCMC for machine learning. Machine learning, 50(1/2): 5–43.

Atkinson P M, Tatnall A R L, 1997. Neural networks in remote sensing: Introduction. *International Journal of Remote Sensing*, 18(4): 699–709.

Batty M, 2013. The New Science of Cities. Cambridge, MA: MIT Press.

Beale C M, Lennon J J, Yearsley J M *et al*., 2010. Regression analysis of spatial data. *Ecology Letters*, 13(2): 246–264.

Benz U C, Hofmann P, Willhauck G *et al*., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(3/4): 239–258.

Brereton R G, Lloyd G R, 2010. Support vector machines for classification and regression. *Analyst*, 135(2): 230–267.

Brunsdon C, Fotheringham A S, Charlton M E, 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4): 281–298.

Brunsdon C, Fotheringham S, Charlton M, 1998. Geographically weighted regression. *Journal of the Royal Statistical Society*: *Series D* (*The Statistician*), 47(3): 431–443.

Byrne G F, Crapper P F, Mayo K K, 1980. Monitoring land-cover change by principal component analysis of multitemporal Landsat data. *Remote Sensing of Environment*, 10(3): 175–184.

Castells M, 1999. Grassrooting the space of flows. *Urban Geography*, 20(4): 294–302.

Castro P S, Zhang D, Li S, 2012. Urban traffic modelling and prediction using large scale taxi GPS traces. Proceeding of Pervasive'12 Proceedings of the 10th International Conference on Pervasive Computing, Newcastle, UK, June 18–22, 2012: 57–72.

Chen J, Ban Y, Li S, 2015. China: Open access to Earth land-cover map. *Nature*, 514(7523): 434.

Chen Long, Stuart Neil, Mackaness A, 2015. Williams. Cluster and hot spot analysis in Lincoln, Nebraska, USA. *Geomatics and Spatial Information Technology*, 38(3): 189–192. (in Chinese)

Chen M S, Han J, Yu P S, 1996. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6): 866–883.

Cheng Changxiu, Shi Peijun, Song Changqing *et al*., 2018. Geographic big-data: A new opportunity for geography complexity study. *Acta Geographica Sinica*, 73(8): 1397–1406. (in Chinese)

Cheng R, Emrich T, Kriegel H P *et al*., 2014. Managing uncertainty in spatial and spatio-temporal data. Proceedings of the IEEE 30th International Conference on Data Engineering (ICDE), Chicago, IL, USA, Mar 31–Apr 04, 2014: 1302–1305.

Congalton R G, 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37(1): 35–46.

Dai J, Yang B, Guo C *et al*., 2015. Personalized route recommendation using big trajectory data. Proceedings of the 2015 IEEE 31st International Conference on Data Engineering (ICDE), Seoul, South Korea, April 13–17, 2015: 543–554.

Data Center of Sina Micro-blog, 2017. 2017 User Development Report of Sina Micro-blog. http://data.weibo.com/report/reportDetail?id=404. (in Chinese)

Du Zhenyu, Xing Shangjun, Song Yumin *et al*., 2007. Lead pollution along expressways and its attenuation by green belts in Shandong province. *Journal of Soil and Water Conservation*, 21(5): 175–179. (in Chinese)

Džeroski S, 2009. Relational Data Mining. Boston, MA: Springer, 887–911.

Ester M, Kriegel H P, Sander J *et al*., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceeding KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, Aug 02–04, 1996: 226–231.

Fayyad U, Piatetsky-Shapiro G, Smyth P, 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11): 27–34.

Friedl M A, Brodley C E, 1997. Decision tree classification of land cover from remotely sensed data. *Remote*

Sensing of Environment, 61(3): 399–409.
Ginsberg J, Mohebbi M H, Patel R S *et al*., 2009. Detecting influenza epidemics using search engine query data. *Natur*e, 457(7232): 1012–1015.
Gislason P O, Benediktsson J A, Sveinsson J R, 2006. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4): 294–300.
Goodchild M F, 2004. The validity and usefulness of laws in geographic information science and geography. *Annals of the Association of American Geographers*, 94(2): 300–303.
Goodchild M F, Yuan M, Cova T J, 2007. Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 21(3): 239–260.
Han J, Lee J G, Kamber M, 2009. An overview of clustering methods in geographic data analysis. *Geographic Data Mining and Knowledge Discovery*, 2: 149–170.
Han Zhigang, Kong Yunfeng, Qin Yaochen, 2011. Research on geographic representation: A review. *Progress in Geography*, 30(2): 141–148. (in Chinese)
Harvey J M, Han J W, 2009. Geographic Data Mining and Knowledge Discovery. London: CRC Press.
Huang Y, Shekhar S, Xiong H, 2004. Discovering colocation patterns from spatial data sets: A general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12): 1472–1485.
Keola S, Andersson M, Hall O, 2015. Monitoring economic development from space: Using nighttime light and land cover data to measure economic growth. *World Development*, 66: 322–334.
Kong X, Xu Z, Shen G *et al*., 2016. Urban traffic congestion estimation and prediction based on floating car trajectory data. *Future Generation Computer Systems*: *The International Journal of EScience*, 61: 97–107.
Koperski K, Han J, 1995. Discovery of spatial association rules in geographic information databases. Proceedings of the 4th International Symposium on Large Spatial Databases (SSD 95), Portland, ME, USA, Aug 06–09, 1995: 47–66.
Koperski K, Han J, Stefanovic N, 1998. An efficient two-step method for classification of spatial data. Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98), Vancouver, BC, Canada, July 11–15, 1998: 45–54.
Lazer D, Kennedy R, King G *et al*., 2014. The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176): 1203–1205.
LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. *Nature*, 521(7553): 436–444.
Li Deren, Cheng Tao, 1995. Knowledge discovery from GIS databases. *Acta Geodaetica et Cartographica Sinica*, 24(1): 37–44. (in Chinese)
Li Deren, Wang Shuliang, Li Deyi *et al*., 2002. Theories and technologies of spatial data mining and knowledge discovery. *Geomatics and Information Science of Wuhan University*, 27(3): 221–233. (in Chinese)
Li Deren, Wang Shuliang, Shi Wenzhong *et al*., 2001. On spatial data mining and knowledge discovery. *Geomatics and Information Science of Wuhan University*, 26(6): 491–499. (in Chinese)
Li X, Yeh A G O, 2002. Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science*, 16(4): 323–343.
Liu Yang, Liu Ronggao, 2015. Retrieval of global long-term leaf area index from LTDR AVHRR and MODIS observations. *Journal of Geo-Information Science*, 17(11): 1304–1312. (in Chinese)
Liu Yu, 2016. Revisiting several basic geographical concepts: A social sensing perspective. *Acta Geographica Sinica*, 71(4): 564–575. (in Chinese)
Liu Z, Ma T, Du Y *et al*., 2018. Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. *Transactions in GIS*, 22(2): 494–513.
Marr B. Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. Chichester, UK: John Wiley & Sons, 2015.
Mayer-Schonberger V, Cukier K, 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. London: John Murray.
McMillen D P, 2004. Geographically weighted regression: The analysis of spatially varying relationships. *American Journal of Agricultural Economics*, 86(2): 554–556.
Miller H J, Goodchild M F, 2015. Data-driven geography. *GeoJournal*, 80(4): 449–461.
Moon T K, 1996. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6): 47–60.
Mutanga O, Adam E, Cho M A, 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation*, 18: 399–406.
NASA, 2017. New Night Lights Maps Open Up Possible Real-Time Applications. https://www.nasa.gov/feature/goddard/2017/new-night-lights-maps-open-up-possible-real-time-applications.

Newing A, Anderson B, Bahaj A B *et al*., 2016. The role of digital trace data in supporting the collection of population statistics: The case for smart metered electricity consumption data. *Population, Space and Place*, 22(8): 849–863.

Niemeijer D, 2002. Developing indicators for environmental policy: Data-driven and theory-driven approaches examined by example. *Environmental Science & Policy*, 5(2): 91–103.

Niu N, Liu X P, Jin H *et al*., 2017. Integrating multi-source big data to infer building functions. *International Journal of Geographical Information Science*, 31(9): 1871–1890.

NOAA/National Centers for Environmental Information, 2018. Global Historical Climate Network Daily: Description. https://www.ncdc.noaa.gov/ghcn-daily-description.

Novembre J, Johnson T, Bryc K *et al*., 2008. Genes mirror geography within Europe. *Nature*, 456(7218): 98–101.

Oliver M A, Webster R, 1990. Kriging: A method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3): 313–332.

Pal M, Mather P M, 2005. Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, 26(5): 1007–1011.

Pei T, Gao J, Ma T *et al*., 2012. Multi-scale decomposition of point process data. *GeoInformatica*, 16(4): 625–652.

Pei T, Sobolevsky S, Ratti C *et al*., 2014. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9): 1988–2007.

Pei Tao, Li Ting, Zhou Chenghu, 2013. Spatiotemporal point process: A new data model, analysis methodology and viewpoint for geoscientific problem. *Journal of Geo-Information Science*, 15(6): 793–800. (in Chinese)

Qian Chengcheng, Cheng Ge, 2018. Big data science for ocean: Present and future. *Bulletin of Chinese Academy of Sciences*, 33(8): 884–891. (in Chinese)

Silver D, Huang A, Maddison C J *et al*., 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587): 484–489.

Silver D, Schrittwieser J, Simonyan K *et al*., 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359.

Song C, Qu Z, Blumm N *et al*., 2010. Limits of predictability in human mobility. *Science*, 327(5968): 1018–1021.

Stein M L, 2012. Interpolation of Spatial Data: Some Theory for Kriging. New York: Springer Science & Business Media.

Sun Zhongwei, Lu Zi, 2005. A geographical perspective to the elementary nature of space of flows. *Geography and Geo-Information Science*, 21(1): 109–112. (in Chinese)

Tobler W R, 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(Suppl.1): 234–240.

Vazifeh M M, Santi P, Resta G *et al*., 2018. Addressing the minimum fleet problem in on-demand urban mobility. *Nature*, 557(7706): 534–538.

Wang Haiqi, Wang Jinfeng, 2005. Research on progress of spatial data mining. *Geography and Geo-Information Science*, (4): 6–10. (in Chinese)

Yuan J, Zheng Y, Xie X *et al*., 2013. T-Drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Transactions on Knowledge & Data Engineerin*g, 25(1): 220–232.

Yuan Y, Wei G, Lu Y, 2018. Evaluating gender representativeness of location-based social media: A case study of Weibo. *Annals of GIS*, 24(3): 163–176.

Zandbergen P A, 2008. Positional accuracy of spatial data: Non-normal distributions and a critique of the national standard for spatial data accuracy. *Transactions in GIS*, 12(1): 103–130.

Zhang Wenjian, 2010. WMO integrated global observing system (WIGOS). *Meteorological Monthly*, 36(3): 1–8. (in Chinese)

Zhao B, Zhang S, 2018. Rethinking spatial data quality: Pokémon go as a case study of location spoofing. *The Professional Geographer*, doi: 10.1080/00330124.2018.1479973.

Zhao Z L, Shaw S L, Xu Y *et al*., 2016. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30(9): 1738–1762.

Zheng Y, Liu Y, Yuan J *et al*., 2011. Urban computing with taxicabs. Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, September 17–21, 2011: 89–98.

Zhu A-X, Qi F, Moore A *et al*., 2010a. Prediction of soil properties using fuzzy membership values. *Geoderma*, 158(3/4): 199–206.

Zhu A-X, Yang L, Li B *et al*., 2010b. Construction of membership functions for predictive soil mapping under fuzzy logic. *Geoderma*, 155(3/4): 164–174.