Considerations on Geospatial Big Data

Zhen LIU, Huadong GUO and Changlin WANG
Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China
E-mail: liuzhen@radi.ac.cn

Abstract. Geospatial data, as a significant portion of big data, has recently gained the full attention of researchers. However, few researchers focus on the evolution of geospatial data and its scientific research methodologies. When entering into the big data era, fully understanding the changing research paradigm associated with geospatial data will definitely benefit future research on big data. In this paper, we look deep into these issues by examining the components and features of geospatial big data, reviewing relevant scientific research methodologies, and examining the evolving pattern of geospatial data in the scope of the four 'science paradigms'. This paper proposes that geospatial big data has significantly shifted the scientific research methodology from 'hypothesis to data' to 'data to questions' and it is important to explore the generality of growing geospatial data ‘from bottom to top’. Particularly, four research areas that mostly reflect data-driven geospatial research are proposed: spatial correlation, spatial analytics, spatial visualization, and scientific knowledge discovery. It is also pointed out that privacy and quality issues of geospatial data may require more attention in the future. Also, some challenges and thoughts are raised for future discussion.

1. Introduction
Currently, big data has been a hot topic worldwide covering academic, governmental and commercial communities. Typically, about 80% of datasets relate to a spatial location [1][2][3]. Scientists usually call this kind of data geospatial data. According to statistics, Google generates 25 PB of data per day, and geospatial data takes up a significant position among it [4]. By 2014, ESA alone had exceeded 1.5 PB of Earth observation data [5].

Researchers have been devoted to complicated technologies, architectures and applications in the big data landscape. Pioneering technologies, such as Apache Hadoop [1] and MapReduce [6], data infrastructures, and analytic tools, have been fully developed. However, few studies have been conducted to look into the evolution of geospatial big data with scientific research methodologies and the capabilities of geospatial data in the overarching (big data) landscape, which are the real driving components for the development of the technologies and architectures mentioned above. Miller and Goodchild [7] argued that we have entered into a data-driven era. It might be more useful and helpful to review and examine the nature of these data than simply rush to advance the relevant technologies in terms of extracting knowledge from huge geospatial data.

By reviewing current research findings, this paper proposes that geospatial big data has been experiencing a huge evolution in terms of its scientific research methodology and the potential rules of

http://hadoop.apache.org
dealing with these geospatial data. Particularly, in the data-driven context, key research focuses are pointed out with regard to the future of research on geospatial big data.

We start by summarizing the components and features of geospatial big data (not general big data). Then, scientific research methodologies of geospatial big data are reviewed to examine the evolving pattern of geospatial data in the scope of the four Science Paradigms. Following that, some visions are discussed in terms of potential key research on geospatial data. At the end, we summarize some challenges and thoughts that might be instructive for future studies.

2. Definition of geospatial big data

Traditionally, geospatial data refers to geo-referenced data that correlates to Earth’s environmental components and processes and further to the interaction between humans and Earth by using spatial technology assisted with ground station systems. The original generation of geospatial data and its outbreak into big data later on greatly benefited from the rapid development of remote sensing, computing, and information communication technologies, among others. It has been noted that the geospatial data increase has been such an explosive trend that it has outpaced the existing capacities and growth rates of storage, computing and analysis systems [8][9]. For example, the amount of remote sensing images produced by advanced airborne, satellite and ground-based remote sensing systems is increasing at the rate of one terabyte per day, and even a single image set can reach tens of gigabytes. Statistically, remote sensing data on a global scale is approaching the exabyte (EB) level (1 EB = 1024 PB, 1 PB = 1024 TB, 1 TB = 1024 GB) [5]. These geo-related, spatiotemporal data are almost always scientifically oriented and mainly controlled by governmental or commercial agencies (although current policies allow much open access to these data), and are normally labeled as ‘authoritative’ or ‘official’ [10][11].

If the explosive growth of spatiotemporal satellite data is attributed to the development of spatial technology, computing, communication, and other relevant technology, then the advancement of social networks, Web 2.0 and mobile devices and the policy of free public access to satellite images has really promoted the collection of public-contributed data [12][13]. During the last decades, some emerging data sources enabled the appearance of other forms of location-related data. These emerging geospatial data encompass in situ sensor network data (e.g., OpenStreetMap), GPS trace data from mobile devices, geo-social media data (e.g., Twitter), and crowdsourcing/VGI data [9][12][14][15][16][17][18]. These emerging geo-related data are mostly contributed by the volunteered public and largely correlate to the creator’s motivation, behavior and circumstance. They are called ‘user-generated’ or ‘volunteered’ data [11][16][19].

With the emergence of ‘user-generated’ data, the concept of geospatial data has been stretched to a broader scope (see the differences in Table 1), and the amount of these data are in turn growing explosively. Along with the increasing amount of data, the analysis and computing technologies and rapid processing capability are accordingly required to constantly improve in order to match the rapid generation of huge geospatial data. In view of this, some researchers define geospatial big data as spatial datasets exceeding the existing capability of computing systems [9].

| Table 1. The definition of geospatial data |
|------------------------------------------|
| **Label** | **Data scope** | **Receiving methods** |
|------------|---------------|----------------------|
| Traditional geospatial data | Authoritative or official data | Geo-referenced spatiotemporal data | Spatial technology, e.g., satellite remote sensing |
| Emerging geospatial data | User-generated or volunteered data | Geo-related behavior-based data | Social networks, Web 2.0 and mobile devices |
3. Features of geospatial big data

Many studies have been conducted on geospatial big data. However, most of them simply point out that geospatial data qualifies as big data because of the ‘3Vs’ (or more) [7][8][9][10][20] with few descriptions of what the Vs of geospatial big data are. In this section, the features of geospatial big data are systematically discussed.

3.1. The Vs

As mentioned, geospatial big data has the features of big data, falling well within the ‘3Vs’, namely volume, velocity and variety [21].

Volume – The images collected from Earth observing satellites contain rich information with high spatial, temporal and radiometric resolution, high acquisition rate, and short observing period. The higher the resolution is, the bigger the size of the image is. Considering that there are more than 500 satellites globally and some satellites have worked for decades (e.g., the Landsat satellites have served almost 30 years), the received satellite image data have become quantitatively enormous [22].

Velocity – The real-time or near real-time monitoring of satellites means a constant flow of image data, which demands high computing and analysis capabilities. Data from new sources, such as VGI, are based on users’ interactivity, so that the value of the data can only be found and used when they are provided, processed and shared dynamically, almost in real-time [7][9].

Variety – (i) Geospatial data has three basic models: raster (grid, e.g., satellites images), vector (encompassing points, lines, and polygons), and graph (spatial network) [17][18]. (ii) Multiple sources of and approaches to data collection, such as the global Earth observation systems based on remote sensing technology and the in situ measurements of sensor networks, mobile devices, geo-related social media, and citizen sensors, have generated various types of geospatial data [22][23][24][25][26]. (iii) Some data are originally derived from sensors or software, but some are generated from complex operational/modeling systems. (iv) Heterogeneous data are produced in various formats, encompassing structured (tables and relations), unstructured (text and imagery), and semi-structured (auxiliary) data [7][22][27].

In addition, other ‘Vs’ have been proposed to define geospatial big data, such as value, veracity, and visualization [20][28].

Value – Scientific technologies have largely advanced to manage and process geospatial data by extracting the essential information (the valuable part) from redundant noise, so as to discover new insights and scientific knowledge [8][29]. However, it is hard to make each dataset valuable because of the huge volume and complexity. As for geospatial data, its big volumes have a very low value density, described as ‘abundant data with scarce information’ [29]. It is impossible and unnecessary to extract all information from huge amounts of data; to find the most valuable data, Turner et al. [30] proposed five criteria to define ‘target-rich’ data. Target-rich data: are easy to access, are real-time, have a large footprint, are transformative, and have ‘intersection synergy’.

Veracity – Data accuracy has always been given attention for various uses in different fields. Satellite images and the processing procedure encompass many uncontrollable factors, and the credibility of social media data is concerning in terms of the measuring accuracy and certainty of the data [31]. The uncertainty of data is a result of it coming from many sources and including noise, deletions, inconsistencies and ambiguities. As Goodchild [32] argued, ‘all location references are subject to uncertainty’, and maybe the veracity issue can only be overcome through modeling and analyzing the huge collection of geospatial data.
Visualization – The term ‘big data’ was originally produced in the context of computer systems being challenged by visualization [33]. The spatial characteristic of geospatial data makes it possible and reasonable to transform and display the data onto the screen to enhance interactive processing with users. With the increase in size and dimension, and the demand for quick display shifting from 2D to 3D presentation, geospatial data also challenges the capacity of computer processing.

3.2. Beyond Vs

Besides the ‘Vs’, the unique characteristics of geospatial data should also be highlighted in terms of space and geography, as it is these characteristics that support the ‘V’ features of geospatial big data and are more important for people to discover the value and relations behind them.

Most basically, geospatial data has the properties of spatial auto-correlation (Tobler’s First Law) and spatial heterogeneity [17]. Spatial auto-correlation means that the attribute values of geographic targets are correlated and more importantly there exists a neighborhood effect. Spatial heterogeneity refers to inconsistent observing results caused by spatial location differences when observing [34]. These features of geospatial big data are much more related to the differences in spatial location of observed targets themselves, rather than the aforementioned variety related to various sources or formats of the observing methods and tools.

Meanwhile, geospatial data have ‘3H’ features: high dimensionality (including spatial, spectral, and temporal dimensions), high complexity (complex modeling and computing methods and systems), and high uncertainty (the uncertain sensing and empirical process) [22][35]. Compared to traditional spatial data, which are mostly static maps and regular or irregular survey data that statically describe Earth’s surface, geospatial big data cover a more refined space and time granularity and are qualified with the typical liquid spatiotemporality. With these characteristics, the emergence of large amounts of data in a short period may lead to unlimited possibilities.

4. Evolving pattern

No matter how big the geospatial data are, the unprecedentedly advanced capabilities of acquiring, storing, processing, computing and analyzing geospatial data make the scarce-data era part of the past in humankind’s understanding of Earth. We have entered into a data-rich era. While some scientists make progress on processing/computing technology, others are starting to think about how people should react to the ‘exaflood’ of data and learn to ‘drink from a fire hose’ [36][37]. This idea stirs up deep thinking about the evolving pattern of geospatial big data. It may be a good time for people to re-evaluate the role of data and make some changes when using data to solve problems.

Data, since the concept originally appeared, have always accompanied the evolution of science and technology. Looking at the four Science Paradigms makes us clearly aware of the relationship between the improvement of science and technology and data [38]. In the Empirical Paradigm, attempts were made to explain natural phenomena—it was a technology-driven phase seeking the accumulation of original data. Moving into the Theoretical Paradigm, models were used to generalize the empirical data—it was a theory-driven phase of verifying data. When entering into the Computational Paradigm, computers were applied to simulate complex phenomena to find any possible rules—it was a model-driven phase of describing and predicting trends. In 2007, the concept of data-intensive scientific discovery was proposed as the Fourth Paradigm—it is a data-driven phase of knowledge discovery (See Table 2).

| Science Paradigm | Driving Pattern | Aims                        |
|------------------|----------------|-----------------------------|
| Empirical        | Technology-driven | Seeking data to explain natural phenomena |

Table 2. The role of data in the four Science Paradigms
The four Science Paradigms show, just as Miller and Goodchild [7] argued, that it is reasonable to describe the development of data as an evolution rather than a revolution. Below we shall look deeply at the evolving pattern of geospatial data from a data-seeking to a data-driven phase. Specifically, the data-intensive paradigm will be examined to find the role of data in knowledge discovery.

- **From hypothesis to data**

  According to classical scientific methodologies, questions or hypotheses (and maybe predictions) were created before executing experiments [39]. Data, which were produced through methods or tools, were used to support or explain what had been queried, hypothesized or predicted. In this sense, the motivation of people seeking data was largely attributed to them having something in mind to test or explain. In this paradigm, the role of data in scientific research is passive.

  Because traditional measuring methods might have been time-consuming, especially prior to the satellite era, the obtained geographic data would have been very limited and scarce in terms of efficiency, scale, and comprehensiveness, compared to that in the big data era. Considering the limitations of data acquired by traditional methods, the usage of data was strongly target-oriented. These ‘sampling’ data were collected to solve certain problems [40], which could lead to the trend of data being typical, not in generalization [7]. Maybe it is necessary to argue about how valid problems are solved by using these data and about the generality of them [7].

- **From data to questions**

  Since the launch of the first Earth observing satellite, *Nimbus*, in 1964, there have been 514 Earth observing satellites as of 2011[5]. Many countries have proposed initiatives for Earth observation based on remote sensing satellite technology. With the establishment of global Earth observing satellite systems and the advanced capability of multi-scale, real-time dynamic monitoring, satellite data has exploded. With these abundant data on hand, users are allowed to examine datasets for themselves to find the useful ones for their particular research purposes.

  As for location data from emerging sources (e.g., social media and mobile devices), they are mostly voluntarily contributed by the public. Any users of social media or mobile devices can be a geo-located data generator and provider. None of them can tell the possible future uses of their contributions in advance. They just generate and share their data through geo-social media or sensor networks individually. However, these individual volunteered location data are compiled into a big data pool and only a set of the data from the pool can possibly assist in solving problems through new technologies, e.g., OpenStreetMap or eco-routing systems.

  It has become a trend that authoritative data and emerging volunteered data might be generated and shared without precise, pre-set hypotheses or problems that need to be tested or to solved. They are quite often collected and stored with a general purpose. For example, some satellite images are aimed at monitoring environmental changes, but to monitor what kind of environmental changes (e.g., water resources, soil use, or vegetation cover), and where, is pendent. Environmentalists and biologists are likely to use data with different spatiotemporal characteristics when conducting their specific research. Taking another example, VGI data is generated and circulated by individuals with few or no exact purposes. But these data may help governmental agencies efficiently make decisions on traffic issues, or assist the public with travel plans.

  Currently, big populations of data are not solely generated for usage in a certain project. Although some thematic and scientific satellites are designed based on pre-set tasks, these data are automatically
received and could be used for other purposes or in other fields. Users are free to access data and to extract the information they are concerned about. In this sense, compared to ‘sampling’ data’s passive mode in scientific research, the ‘population’ property of geo-related data is more active and powerful in driving people to discover knowledge [40].

- **Challenges**
  
  The shift from ‘hypothesis to data’ to ‘data to questions’ coincides so well with geographical thinking that it may be much more reasonable to explore the world in a ‘bottom to top’ view [41]. Instead of talking about how to collect data to solve existing problems, people increasingly think about what kinds of problems could be addressed by using the collected data. This helps to find the generality of big data, and also leads to its broader usage in the future.

  Normally, with excessive focus on computing capacity, models, and algorithms in specific fields, scientists pay little attention to the generality of explosively growing geospatial data. This can result in scientific hypotheses or questions being left to those specific scientists to think about and solve. With the increasing accumulation of scientific hypotheses and questions, the common scientific questions (or generality) of big data can be extracted and proposed in hopes of overcoming scientific barriers [42].

  When faced with big spatial data, it is easy to imagine a scenario where researchers are overwhelmed with choices, like a hungry man at a buffet. How do they select the right data from the big data pool either to address a certain problem or to make a decision? How do they acquire effective information or knowledge from such a big data pool? Maybe these are issues that only data scientists can address. Maybe we need to establish auxiliary information classification or filtering mechanisms to assist with information retrieval and knowledge discovery.

5. **Future research on geospatial big data**

In the scarce-data era, the use of computers made processing and analyzing data convenient, and somehow needed a growing amount of data to satisfy and verify advancements. This changed dramatically when entering into the big data era due to the ability to collect data no longer being a problem. The computer loses control of the data to some extent when data comes like an ‘exaflood’ [36][37]. Moreover, scientific and technical methodologies strongly improved to meet the demands caused by the growing amount of data. No longer hindered by the capabilities of collecting and computing, geospatial data enabled people to exploit and discover knowledge. In this sense, big data became a driving force, and plays a pivotal role together with technologies in helping human beings meet various challenges.

Undoubtedly, we have entered into a data-driven phase in geographic research. Given the circumstance that data collection and relevant processing technologies are not the terminal goal of big data research, how to effectively use these data has become the core of attention. In this section, four potential research areas for geospatial data are proposed.

- **Spatial Correlation**
  
  It has been discussed that the correlation among data is more valuable in scientific knowledge discovery than the data themselves or even models that describe these data [35]. Differing from traditional logical inference, it is more suitable to employ analytic induction to find correlation or relevance in huge volumes of data by using searching, comparing, clustering, and classifying technologies. Through correlation analytics, association networks that might be hidden in the data can be figured out [42].

  In terms of geospatial big data, it is argued that research on geographic science has developed beyond the phase of explaining phenomena or findings (knowing why), and rather focuses more on
correlations of the data collected from in situ experiments, sensor monitoring, computer generation, and even individual observations. It is the correlation that enables geographic scientists to produce the detection and representation patterns of data and to make possible predictions [7]. As discussed before, more than 80% of data are related to location [1][2][3], making it meaningful to examine the spatial correlation of geospatial data based on their location information.

- **Spatial Analytics**
  Traditional spatial analytics refers to using appropriate statistical analysis and artificial intelligence algorithms to analyze geospatial data, extract useful information, and summarize the general process. With the increase in magnitude of geospatial data, traditional methods for spatiotemporal analysis of geospatial big data have been unable to meet demands, and must be improved. Spatial partitioning, multi-dimensional data structures, static and dynamic load balancing, multiple iterations, and modeling algorithms should be taken into account [17].

Given that geospatial data consists of both location and time information, integrating and analyzing timely data (e.g., VGI data) makes it possible to detect the data generator’s mobility habits and model their routes. This capability benefits greatly from advanced trajectory modeling technologies [43][44][45][46]. Along with this behavioral information, geospatial data also contain a wealth of other social information. Some researchers argue that existing models cannot meet the requirements of processing non-spatiotemporal data that accompany the raw geospatial data [47]. With the involvement of social network activities, human behavioral patterns and the context within which data are generated, high-dimensional spatial analysis models are needed. Such advanced models may be complex, but could be more precise in representing the reality of the data and help people extract useful information and form basic knowledge from the raw geospatial data.

- **Spatial Visualization**
  Geospatial data is characterized by high dimensionality, including high spatial, spectral and temporal dimensions. With the continuous development of computing graphics, imaging techniques and data acquisition techniques (e.g., LiDAR), the ability to collect spatial-topological information has been gradually strengthened. The increase in size and dimension of geospatial data demands the enhancement of spatial data visualization. Although 3D and 4D technology has been developed, traditional representation and visualization methods are still largely restricted by the two-dimensional display of computers. In addition to computing technology, key technologies, such as multi-dimensional spatial data models, spatiotemporal integration management technology, and 3D spatiotemporal integration modeling methods, are needed to visualize geospatial big data [48].

- **Scientific Knowledge Discovery**
  The fourth paradigm of data-intensive scientific discovery raises new concerns about analyzing and mining big data in terms of relevant correlation and rules in order to find, new knowledge and even new rules that the previous scientific methodologies could not discover [35]. As for geographic knowledge, it was initially derived from ground-based observations and measurements. These data are limited mostly at temporal and spatial scales. Geospatial data, achieved by space-air-ground monitoring technology and systems, provides long-term, multi-scale (from local, regional to global), study area-oriented data for geographic research. Based on rich, multi-source data, advanced computing and analysis technology entail the task of value exploration.

From big data, to data-intensive scientific data, and then to geospatial big data, a variety of models and algorithms have been generated for analysis and interpretation. But rather than only practicing computing capability, the ultimate target of acquiring data is to describe reality, discover knowledge, support decision making, and understand the real world.

6. Challenges
Although the collection of massive geo-related data provides unprecedented opportunities for human beings with abundant information and knowledge to understand Earth, we have to admit that geospatial big data also brings challenges for privacy and data quality issues.

As discussed in the Introduction, authoritative geospatial data are largely controlled by governmental and commercial agencies. The large-scale commercial mode of utilizing these geospatial data runs the risk of reducing necessary, proper applications. In spite of the increasingly open policies for access to satellite data and free mechanisms for sharing data to reduce such risks to some extent, there still exists a challenge caused by open data policies in terms of privacy. Furthermore, because data are contributed from crowdsources or individual volunteers, the reproducibility and privacy of data also create challenges [17].

In addition, multi-source geospatial data raise a critical problem with data quality, which turns out to be a prominent challenge when analyzing and using these data [49]. The massive Earth observation data from aerial and satellite remote sensing technology are quite varied in terms of their technical parameters, storage formats, image resolutions, and observation scales (in both space and time). This leads to some problems in using multi-source heterogeneous geospatial data. Similarly, user-generated geospatial data, such as VGI data, are based on public input. They are timely, but challenge the balance between efficiency and quality. At the same time, public-contributed data are a necessary supplement to scientific data. The availability of such timely, low-cost data brings significant changes to research in social sciences as well as natural sciences [50][51]. However, considering the identities of data contributors and the environment in which data are collected, the data become suspicious in terms of authenticity, credibility and reliability [52][53][54]. Taking OpenStreetMap as an example, discussions have continued about the position accuracy, completeness, and currency [55][56][57][58][59][60].

7. Conclusion
Geospatial data, as an important portion of big data, has gained widespread attention. Depending on the different collection sources and methods, geospatial data can be defined in different scopes. Aside from the ‘3Vs’ (and other ‘Vs’), geospatial big data has its own unique features.

In the big data era, data processing and analysis technologies have been strongly driven by massive data. Compared to the acquisition of data, the capability of extracting information and discovering knowledge from data seems to have greater requirements. Accordingly, the role of geospatial big data in scientific methodology has changed from being produced to test hypotheses to being exploited to discover knowledge, i.e., from ‘hypothesis to data’ to ‘data to questions’.

We pointed out four main future research areas of geospatial data, covering spatial correlation, analytics, visualization and scientific knowledge discovery. Finally, there are privacy and quality issues caused by geospatial big data and a new set of challenges to meet in the future.

8. References

[1] Densham P J and Goodchild M F 1989 Spatial decision support systems: a research agenda Proc. GIS /LIS’89 (Orlando FL)

[2] Shekar S and Xiong H 2007 Encyclopedia of GIS (New York: Springer)

[3] Li Q and Li D 2014 Big Data GIS Geomatics and Information Science of Wuhan University 39 6

[4] Vatsavai R R, Ganguly A, Chandola V, Stefanidis A, Klasky S, Shekhar S 2012 Spatiotemporal data mining in the era of big spatial data: algorithms and applications Proc. 2nd ACM
He G J et al 2015 Processing of earth observation big data: challenges and countermeasures Chin. Sci. Bull. 60 pp 470–478

Dean J and Ghemawat S 2008 MapReduce: simplified data processing on large clusters Commun. ACM 51 1 pp 107-113

Miller H J and Goodchild M F 2014 Data-driven geography GeoJournal

Xu C and Yang C 2014 Introduction to big geospatial data research Ann. GIS 20 4 pp 227-232

Lee J G and Kang M 2015 Geospatial big data: challenges and opportunities Big Data Res. http://dx.doi.org/10.1016/j.bdr.2015.01.003

Schade S 2015 Big data breaking barriers – first steps on a long trail Proc. 36th Int. Symp. on Remote Sensing of Environment (Berlin, Germany)

Bakillah M, Lauer J, Liang S H L, Zipf A, Arsanjani J J, Mobasher A and Loos L 2014 Exploiting big VGI to improve routing and navigation services Big Data Techniques and Technologies in Geoinformatics ed. H A Karimi (CRC Press)

Heipke C 2010 Crowdsourcing geospatial data ISPRS J. Photogramm. Remote Sens. 65 pp 550–570

Salk C, Sturn T, See L, Fritz S and Perger C 2016 Assessing quality of volunteer crowdsourcing contributions: lessons from the cropland capture game Int. J. Digital Earth 9 4 pp 410–426

Ramm F, Topf J and Chilton S 2011 OpenstreetMap: using and enhancing the free map of the world (England: UIT Cambridge)

Howe J 2006 The rise of crowdsourcing Wired Magazine 14 6

Goodchild M F 2007 Citizens as sensors: the world of volunteered geography GeoJournal 69 4 pp 211–221

Cugler D C, Oliver D, Evans M R, Shekhar E S and Medeiros C B 2015 Spatial big data: platforms, analytics, and science GeoJournal

Evans M R, Oliver D, Zhou X and Shekhar S 2014 Spatial Big Data Case Studies on Volume, Velocity, and Variety Big Data Techniques and Technologies in Geoinformatics ed H A Karimi (CRC Press)

Coleman D, Georgiadou Y and Labonté J 2009 Volunteered geographic information: the nature and motivation of producers Int. J. Spatial Data Infrastructures Res. 4 pp 332–358

Nativi S, Mazzetti P, Santoro M, Papeschi F, Craglia M and Ochiai O 2015 Big Data challenges in building the Global Earth Observation System of Systems Environmental Modelling & Software 68 pp 1–26
[21] Laney D 2001 3D Data Management: Controlling Data Volume, Velocity and Variety (Gartner) Available at: http://blogs.gartner.com/doug-laney/files/

[22] Song W, Liu P, Wang L and Lv K 2014 Intelligent Processing of Remote Sensing Big Data: Status and Challenges, *J. Eng. Stud.* 6 3 pp 259–265

[23] Miller H J 2007 Place-based versus people-based geographic information science *Geography Compass* 1 3 pp 503–535

[24] Miller H J 2010 The data avalanche is here. Shouldn’t we be digging? *J. Regional Sci.* 50 1 pp 181–201

[25] Sui D and Goodchild M F 2011 The convergence of GIS and social media: Challenges for GIScience *Int. J. Geographical Inf. Sci.* 25 11 pp 1737–1748

[26] Townsend A 2013 *Smart cities: Big data, civic hackers, and the quest for a new utopia* (New York: Norton)

[27] Dumbill E 2012 What is big data? An introduction to the big data landscape, [http://strata.oreilly.com/2012/01/what-isbig-data.html](http://strata.oreilly.com/2012/01/what-isbig-data.html)

[28] Fromm H and Bloehdorn S 2014 Big Data-Technologies and Potential *Enterprise-Integration* ed G Schuh and V Stich (Berlin: Springer) pp 107–124

[29] Tao X, Hu X and Liu Y 2013 Review on Big Data Research *J. System Simulation* 25 (Supplement) pp142–146.

[30] Turner V, Gantz J F, Reinsel D and Minton S 2014 The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things *IDC Analyze the Future* [http://www.emc.com/leadership/digital-universe/2014iview/high-value-data.htm](http://www.emc.com/leadership/digital-universe/2014iview/high-value-data.htm)

[31] Castillo C, Mendoza M and Poblete B 2011 Information Credibility on Twitter *WWW 2011, Hyderabad*, March 28–April 1

[32] Goodchild M F 2010 Twenty years of progress: GIScience in 2010 *J. Spatial Inf. Sci.* 1 pp 3–20

[33] Cox M and Ellsworth D 1997 Application-controlled demand paging for out-of-core visualization *Proc. IEEE 8th Conf. on Visualization*

[34] Shekhar S, Evans M R, Kang J M and Mohan P 2011 Identifying patterns in spatial information: A survey of methods *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 3 pp 193–214

[35] Guo H D *et al* 2014 Scientific big data and digital Earth *Chin. Sci. Bull.* 59 pp 1047–1054

[36] Sui D, Goodchild M F and Elwood S 2013 Volunteered geographic information, the exaflood, and the growing digital divide *Crowdsourcing geographic knowledge* ed D Sui, S Elwood and M F Goodchild (New York: Springer) pp 1–12

[37] Waldrop M M 1990 Learning to drink from a fire hose *Science* 248 4956 pp 674–675
[38] Gray J 2009 eScience: a transformed scientific method The fourth paradigm: Data-intensive scientific discovery ed T Hey, S Tansley and K Tolle (Microsoft Research)

[39] Gauch H G 2003 Scientific method in practice (Cambridge University Press)

[40] Mayer-Schonberger V and Cukier K 2013 Big Data (Houghton Mifflin Harcourt Publishing Company)

[41] Goodchild M F 2012 Invigorating GIScience www.geog.ucsb.edu/~good/presentations/uugismay12.pptx

[42] Li G and Cheng X 2012 Big data research: the significant strategic areas of the future development of the science, technology, economy and society - the research status of big data and scientific thinking Proc. Chin. Acad. Sci. 27 pp 647–657

[43] Ashbrook D and Starner T 2003 Using GPS to learn significant locations and predict movement across multiple users Personal and Ubiquitous Computing 7 5 pp 275–286

[44] Quannan L, Zheng Y, Xing X, Yukun C, Wenyu L and Ma W Y 2008 Mining user similarity based on location history Proc.17th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems 34 1–10

[45] Giannotti F, Nanni M, Pinelli F and Pedreschi D 2007 Trajectory pattern mining Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (San Jose, CA) pp 330–339

[46] Yu Z, Zhang L Z, Xie X and Ma W Y 2009 Mining correlation between locations using human location history Proc.17th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems (Seattle, WA) pp 472–475

[47] Assam R and Seidl T 2014 Insights and Knowledge Discovery from Big Geospatial Data Using TMC-Pattern Big Data Techniques and Technologies in Geoinformatics ed H A Karimi (CRC Press)

[48] Li J 2014. Research on key techniques of 3D visualization of complex spatial data J. Scientific and Technical Entrepreneurship

[49] Goodchild M F 2013 The Quality of Big (Geo) Data Dialogues in Human Geography 3 3 pp 280–284

[50] King G 2011 Ensuring the Data-Rich Future of the Social Sciences Science 331 pp 719–721

[51] Goodchild M F and Li L 2012 Assuring the quality of volunteered geographic information Spatial Statistics 1 pp 110–120

[52] Elwood S, Goodchild M F and Sui D 2013 Prospects for VGI Research and the Emerging Fourth Paradigm Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice ed D Sui, S Elwood and M F Goodchild (New York: Springer) pp 361–375

[53] Tang J C, Cebrian M, Giacobe N A, Kim H W, Kim T and Wickert D B 2011 Reflecting on the Darpa Red Balloon Challenge Communications of the ACM 54 4 pp 78–85
[54] Flanagin A J and Metzger M J 2008 The credibility of volunteered geographic information, *GeoJournal* 72 pp 137–148

[55] Haklay M 2010 How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets *Environment and Planning B: Planning and Design* 37 pp 682–703

[56] Haklay M, Basiouka S, Antoniou V and Ather A 2010 How many volunteers does it take to map an area well? The validity of Linus’ Law to volunteered geographic information *The Cartographic J.* 47 4 pp 315–322

[57] Neis P, Zielstra D and Zipf A 2011 The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011 *Future Internet* 4 4 pp 1–21

[58] Canavosio-Zuzelski R, Agouris P and Doucette P 2013 A photogrammetric approach for assessing positional accuracy of OpenStreetMap roads *ISPRS Int. J. Geo-Information* 2 2 pp 276–301

[59] Hecht R, Kunze C and Hahmann S 2013 Measuring completeness of building footprints in OpenStreetMap over space and time *ISPRS Int. J. Geo-Information* 2 4 pp 1066–1091

[60] Asahara A, Maruyama K, Sato A and Seto K 2011 Pedestrian-movement prediction based on mixed Markov-chain model *Proc. 19th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems* (Chicago, IL) pp 25–33

**Acknowledgments**

This study is supported by the Open Foundation of the Key Laboratory of Digital Earth Sciences, Chinese Academy of Sciences (project number 2014LDE001).