ISPRS Journal of Photogrammetry and Remote Sensing xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Review Article

Rethinking big data: A review on the data quality and usage issues

Jianzheng Liu [a], Jie Li [a], Weifeng Li [a,*], Jiansheng Wu [b,c]

[a] Department of Urban Planning and Design, Faculty of Architecture, Knowles Building, The University of Hong Kong, Pokfulam Road, Hong Kong
[b] Key Laboratory of Human Environmental Science and Technology, Room E318, Peking University Shenzhen Graduate School, University Town, Shenzhen 518055, China
[c] Key Laboratory for Earth Surface Processes, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 18 May 2015
Received in revised form 17 November 2015
Accepted 17 November 2015
Available online xxxx

Keywords:
Big data
Data quality and error
Data ethnics
Spatial information sciences

ABSTRACT

The recent explosive publications of big data studies have well documented the rise of big data and its ongoing prevalence. Different types of "big data" have emerged and have greatly enriched spatial information sciences and related fields in terms of breadth and granularity. Studies that were difficult to conduct in the past time due to data availability can now be carried out. However, big data brings lots of "big errors" in data quality and data usage, which cannot be used as a substitute for sound research design and solid theories. We indicated and summarized the problems faced by current big data studies with regard to data collection, processing and analysis: inauthentic data collection, information incompleteness and noise of big data, unrepresentativeness, consistency and reliability, and ethical issues. Cases of empirical studies are provided as evidences for each problem. We propose that big data research should closely follow good scientific practice to provide reliable and scientific "stories", as well as explore and develop techniques and methods to mitigate or rectify those 'big-errors' brought by big data.

1. Introduction

The prevalence of big data exerts a profound impact on many disciplines, including public health and economics (Einav and Levin, 2014; Khoury and Ioannidis, 2014). Almost all disciplines and research areas, including computer science, business, and medicine, are currently deeply involved in this spreading computational culture of big data because of its broad reach of influence and potential within multiple disciplines (Boyd and Crawford, 2012). As a highly interdisciplinary subject, space information science and related disciplines (e.g., geography and urban studies) are also largely affected by the new technical wave of big data. The past several years have seen the popular applications of big data, such as inferring people's daily travel behavior and interaction using mobile phone data and taxi trajectory data. We can foresee that the wave of big data will eventually be extended to other city applications such as real-time population census and energy use at home or in vehicles. The key question is no longer technological but organizational (Batty, 2012). However, "big data is part of the wave but that is just data. Data only matters if it is useful" (Webster, 2014). We argue that big data also brings problems in data quality and data usage, which undermine the usability of big data. Research based on data with errors don't meet the requirements of good scientific research in terms of authenticity and accuracy. This type of research will likely result in biased or wrong conclusions if we do not have a deeper understanding of the quality issues of big data and its consequent problems.

This study reviews existing literature on big data in spatial information sciences and related fields to obtain an understanding of the current hype on big data and its data quality. We attempt to determine the typical data quality and data usage problems that undermine the authenticity and reliability of big data research in this field. Our intention is not to discourage big data research but to promote a scientific and reliable research culture for big data studies and to facilitate the production of high-quality research.

This review is comprised of three sections. We first present an overview of big data research in spatial information sciences and related fields. This section clarifies the definition of big data and summarizes the current application scope of big data in spatial information sciences and related fields. Several influential empirical publications that focus on big data are highlighted. We also explicate the three paths wherein big data influence spatial information sciences and related disciplines which are data collection, data processing and data analysis; and we assess current big data

* Corresponding author.
E-mail addresses: jzliu@hku.hk (J. Liu), Jessieleepku@hotmail.com (J. Li), wfli@hku.hk (W. Li), wujs@pkusz.edu.cn (J. Wu).

http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006
0924-2716/

studies from the perspective of the three paths respectively. We then focus on the 'big errors' in data collection, processing, and analysis for big data in spatial information sciences and related fields. We elaborate on the five data quality and data usage issues of big data, namely, authoritativeness problem, information incompleteness and noise problem, representativeness problem, consistency and reliability issues, and ethical problems. Cases of empirical studies are presented as evidence. Finally, this paper presents several specific coping strategies and recommendations to help decrease the data error of big data research.

2. Overview on big data research in spatial information sciences

2.1. What exactly is big data and what makes them popular

Big data is generally considered linkable information that have large data volumes and complex data structures (Khoury and Ioannidis, 2014), such as social media data, mobile phone call records, commercial website data (e.g., eBay, Taobao), volunteering geographical information, search engine data, smart card data, and taxi trajectory data. Big data came into the focus of academics only in the past decade as shown in Fig. 1, but the explosive publications of big data studies show that big data topics will probably continue to proliferate in the next few years.

The most popular description of big data thus far is the "3V" model, where "3V" refers to volume, variety, and velocity (Laney, 2001). Volume literally means that typical big data have particularly large data volume. For example, mobile phone call records usually have 70 million data entries (Gao et al., 2013), video surveillance records can even have larger data volume in terms of data storage. Variety means that big data have diversified data sources, data structures, and potential applications. Velocity refers to the real time or quasi-real-time data updating. For instance, air quality monitoring data are often updated once or several times each day. In addition to the "3V" model, "4V" and "5V" models are emerging as researchers attempt to redefine big data. IBM promotes the conformance to veracity to explain the bias problems brought by big data and believes that the "4V" model can accurately describe big data (IBM, 2013). Several media columns argue that big data also have the features of value, variability, and visualization (McNulty, 2014).

However, the typical "big data" in spatial information sciences and related fields appears unfit for the "4V" big data model. Some "big data" such as the social media data of a specific topic are small in terms of data volume and are even smaller than some traditional datasets such as census data. Big data is more about the capacity to search, aggregate, and cross-reference large datasets than its large volume (Boyd and Crawford, 2012). Thus, we argue that "fine-scale spatial–temporal data" will be a more appropriate term to describe the big data in spatial information sciences and related fields since the big data in these fields is usually characterized by a very fine granularity and spatial–temporal dimensions.

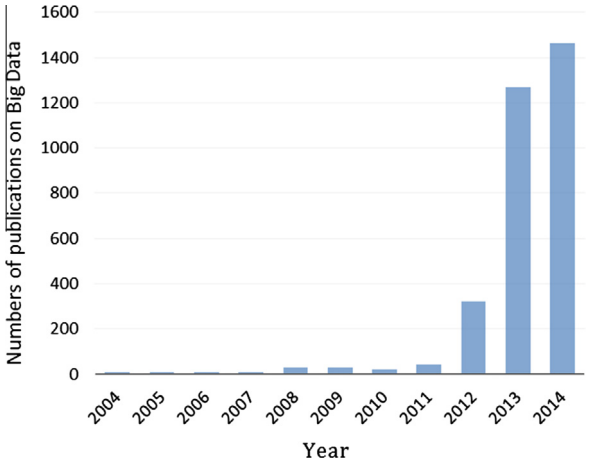


Fig. 1. Number of published studies on "big data" based on literature search of the phrase "big data" in the topic field in the database of web of knowledge from the year 1956 to 2014.

We believe that one of the reasons why big data is popular in most disciplines is that it largely improves the data availability and accessibility of research subjects, thus allowing the study of topics that were difficult to interrogate because of poor data availability. Big data provide "the capacity to collect and analyze data with an unprecedented breadth and depth and scale" (Lazer et al., 2009). For example, obtaining detailed data on the spatial–temporal behavior of urban residents used to be difficult. However, such information has now become accessible and easy to collect because of the popularity of personal communication devices with smart sensors.

Another important feature of big data that makes it prevalent is that it provides extraordinary fine-grained detailed data in terms of analysis units, spatial, and temporal resolution. For instance, smart card and mobile phone data are collected at the individual level (Richardson et al., 2013). Such data can be observed at short intervals, for example, on a per-hour basis. Data with fine analysis units offer a significant chance for rigorous and accurate research because researchers can examine the causal relationship in a small analysis unit and avoid ecological fallacy and the other issues caused by data aggregation (Robinson, 2009). Furthermore, the fine spatial and temporal resolution of big data enable researchers to look into urban issues and other geographical processes in fine detail to generate new understanding and theories, because most current theories are built on radical and massive changes to urban issues and other geographical processes instead of gradual and subtle changes which are probably more important (Batty, 2012).

2.2. A glimpse of big-data-related research

Big data research basically is data driven in almost every discipline and field. Therefore, big data research either focuses on methodological innovation or prioritizes the application of big data on different topics in geography and urban studies. The scope of big data research is difficult to summarize because big data have different types and each type has different applications. Methodological big data studies are generally computation intensive. For example, a few scholars have proposed innovative computational framework for data mining on big data (Gao et al., 2014; Wu et al., 2014). Notable studies include urban computing (Zheng et al., 2013) and the application of machine learning techniques such as neural network and deep learning to big data analysis (O'Leary, 2013; Pijanowski et al., 2014). Visualization tools and techniques are also becoming popular (Cheshire and Batty, 2012).

The most frequently investigated topics in the application of big data in this field is human mobility (Gao et al., 2013; Gonzalez et al., 2008; Liu et al., 2012; Pei et al., 2014; Roth et al., 2011; Song et al., 2010), followed by spatial interaction (Gao et al., 2013; Krings et al., 2009) and urban structure patterns (Lee et al., 2013; Toole et al., 2012; Yuan et al., 2012).

Significant progress has been achieved in big data research in spatial information sciences and related fields. Table 1 shows several empirical studies in the spatial information sciences and related fields. These studies are selected based on their potential research impact, diversified big data types and research problems. We highlight these studies by summarizing the study focus, data source, methods, and results for each empirical study. This

Please cite this article in press as: Liu, J., et al. Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006

Table 1
Summary of selected big data research in spatial information sciences.

Entry	Study focus	Big data	Method	Result
Gonzalez et al. (2008)	Human mobility	Mobile phone data	Statistical fitting	Human trajectories show a high degree of temporal and spatial regularity.
Roth et al. (2011)	Human mobility	London subway "Oyster" card data	Null model	A polycentric structure is composed of large flows organized around a limited number of activity centers.
Krings et al. (2009)	Spatial interaction	Mobile phone data	Gravity model	Communication intensity between two cities is proportional to the product of their sizes divided by the square of their distance.
Huang et al. (2015)	Human mobility	GPS Trajectories	Markov chain model with consideration of activity changes	The proposed method improves the accuracy in analyzing and predicting human movement.
Zheng et al. (2013)	Urban computing	Air monitoring data and points of interest	A semi-supervised learning approach based on the artificial neural network and conditional random field	The proposed method has advantage over decision tree, conditional random field, and artificial neural network.
Fu and Chau (2013)	Data quality	Social media (Sina Weibo)	Random sampling approach	Representative and reliable statistics on Chinese micro-bloggers are limited.
Haklay (2010)	Data quality	OpenStreetMap	Comparison with Ordinance Survey datasets	OpenStreetMap information can be fairly accurate in terms of positional accuracy.

Table 2
Research contribution of different research scenarios.

	New phenomenon/problem/topic	New method	New data/context
Old phenomenon/problem/topic	Nil	Good (methodological study)	Problematic if without new insights
Old method	Good (new areas)	Nil	Problematic if without new insights
Old data/context	Good (new areas)	Good	Nil

summary can serve as a quick overview of the achievements of big data research in this field.

2.3. How does big data change current research

We believe that big data exert their impact on spatial information sciences and related fields in three aspects, namely, data collection, data processing, and data analysis.

First, the data collection approach has been transformed from traditional methods (e.g., questionnaires and interviews) into a fast and powerful ICT-based method, including the web service provided by different data vendors such as national environmental protection agencies and commercial institutions, as well as devices with sensors such as mobile phones and smart transportation cards. This transformation is fundamental because it has changed the other two aspects, namely, the way researchers process and analyze data.

The change in data collection has led to changes in data processing. Big data is characterized by high volume, velocity, variety, veracity, fine granularity, and rich data availability. Consequently, the methods and procedures to process these big data must have the capability to handle high volume and real-time data and serve as a filter to decrease data errors and data "noise." The third aspect of changes brought by big data in current research is data analysis. The data structure and fine granularity of big data requires new data analysis methods and tools because many existing tools and instruments for data analysis are for the traditional data of coarse temporal and spatial resolutions (Cheshire and Batty, 2012).

2.4. Evaluating current big data research

The emergence of big data has clearly been transforming the research landscape. The transition of data collection has increased the richness and availability of data. A boom in the research areas is expected, as evidenced by the extensive publications in recent years (Fig. 1). However, it appears that some researchers are immersed in this carnival of big data and overlooked its potential problems. These researchers tend to embrace big data without any doubt and scrutiny.

The process of data collection and processing for big data is expected to play a key role in avoiding the systematic biases of big data and decrease the data "noise" to ensure the appropriate use of big data in authentic scientific inquiries. However, we have not seen this happen, as shown in the "big errors" section that follows in this review.

Data analysis is expected to change in this new era of big data. The feature of big data requires new approaches and tools that can accommodate big data with different data structures and can process data with different spatial and temporal scales. However, thus far we have not found any data analysis method that is significantly different from the traditional approach. Some of the existing big data studies basically follow a research paradigm of combining "new approaches based on new data" with old topics (Yuan et al., 2012). In attempt to explore the big data, these studies usually apply or develop methods based on traditional data mining techniques, and use this seemingly 'new' approach to explore an old topic. Original contribution for an empirical research usually comes from either exploring a new topic/phenomenon, or using a new method/model, or gaining new insights. Only employing new data is not enough for a good and original empirical study, as shown in Table 2. Some of the current big data studies run the risk of merely doing data exercise instead of making original contribution. Whether these studies contribute to the understanding of research problems, produce any new insights into the research problem, or improve the theories that explain the real world is doubtful.

3. "Big-errors brought by big data

Anderson (2008), former editor-in-chief of *Wired* magazine, indicated the following in his article "The end of theory": "out with

Please cite this article in press as: Liu, J., et al. Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006

every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves."

However, does big data really have "unprecedented fidelity" and speak for themselves without a theory? We doubt these ideas. We believe that the bigness of big data not only refers to its large data volume, complex data structure, and fine granularity but also to the significance of data quality and usage problems in big data. Researchers in the field of public health, biology and information communication technology have similar view with us on the claim by the Wired magazine (Boyd and Crawford, 2012; Khoury and Ioannidis, 2014; Pigliucci, 2009). This review examines current empirical studies on big data and summarizes the prevailing problems during data collection, processing, and analysis of big data to elucidate the "big errors" of big data studies in spatial information sciences and related fields. We attempt to provide a constructive understanding and argument for the reflection and reexamination of the data authenticity problem of big data studies, as well as suggest concrete measures to mitigate these problems. Given our limited experience on big data, this study focuses on mobile phone data, social media, volunteering data, and searching engine data. Thus, each problem discussed in this review may not be applicable to all types of big data and each type of big data may suffer from one or multiple problems in different ways.

3.1. Inauthentic data collection

Traditional data collection is usually conducted or supervised by scientific investigators, research institutions, or governmental agencies. The data collected by these authoritative scientific bodies generally have high data authenticity and credibility because researchers in these organizations generally obey research ethics and follow good scientific practices. These institutions also have more resources and power to perform these tasks. Furthermore, these scientific data collection tasks are what these researchers, research institutions, and governmental agencies are hired and paid to do. It is their job.

However, some of the big data suffer from authenticity and credibility problems in data collection. For example, social media data are collected from Twitter, Sina Weibo, and other social networking platforms. These organizations are commercial companies that are not established for scientific research purposes but are business platforms that pursue profits. At least three types of differences between a scientific research institution and a commercial company exist. First, commercial business platforms neither adopt scientific data collection procedures such as random sampling method for data collection nor follow a series of solid and scientific data processing procedures to address biases and other data problems. Commercial business platforms perform data collection not for the sake of science but for profit. The collected data are "repurposed" data that have been previously used for commercial purposes but is now used for scientific purposes (Loshin, 2012). The target populations of these companies are people who are profitable to them rather than the overall population. Furthermore, the sampling method and processing algorithms behind the web service provided by these companies are unknown, just like a black box. Second, these commercial big data providers can change the sampling methods and processing algorithms at any time without any notice. Researchers may not know these changes at all. Research that adopts these data generated by different algorithms for temporal comparison can produce biased or even wrong conclusions. Third, commercial platforms have no obligation or motivation to ensure the authenticity and validity of the data they collected. A good example is that social media such as Twitter and Sina Weibo have many robots or "zombie" accounts that are run by machines. These companies have no desire to "eliminate" these unqualified samples. Instead, these companies count on these accounts to make money. Other big-data-based web services, such as Facebook, Fang, Baidu, and Google, have similar problems.

The reputable Google Flu Trends Index is an example of inauthentic data collection. Google reports that this index can use the search records of users who search influenza on Google to predict the breakout of an influenza pandemic. This index has attracted considerable attention, but scholars have reported that Google often changes the algorithms and make the prediction unstable (Lazer et al., 2014). The comparison between the Google estimates of influenza and the influenza records of the US Centers for Disease Control and Prevention (CDC) shows that Google estimates of doctor visits for influenza-like illness are more than double those of CDC estimates, and Google estimates are also higher than those of the CDC in 100 out of 108 weeks (Lazer et al., 2014).

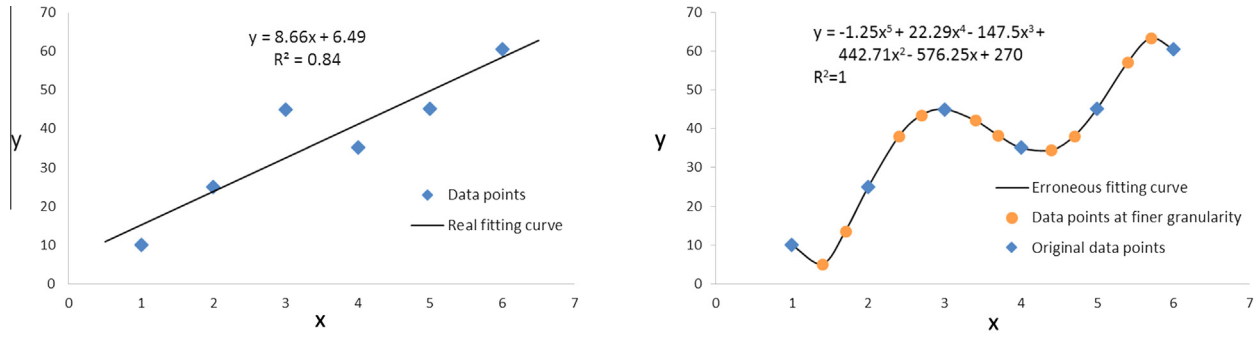
3.2. Information incompleteness and noise of big data

As discussed in Section 3.1, some big data are repurposed from commercial use to scientific use, thus resulting in many data problems. Information incompleteness is one of these problems. Some big data are good in data volume but contains limited information. This situation constrains the further application of these data. For example, mobile phone call detail record (CDR) data records the calling log of many phone users in a city in a real-time manner at a low cost. However, the information in the data is incomplete and has a narrow application scope without the socio-economic attribute data of users.

This condition can be attributed to the limited number of data fields of mobile phone data, including only pseudo user IDs, base station locations, and calling time stamps. Although the data is recorded at the individual level, mobile phone data have extremely limited applications because of the lack of socio-economic attributes. This type of data cannot reflect the differences among respondents, or describe residents' behavior characteristics and other essential information that interests researchers and readers. Current literature shows that mobile phone data can only be used to look into the spatial aspects of human activities and not the socio-economic aspects, which either interrogates human mobility (Gonzalez et al., 2008; Pei et al., 2014) or examines the spatial structure and interaction between or within cities (Gao et al., 2013; Krings et al., 2009; Reades et al., 2009; Soto and Frías-Martínez, 2011; Toole et al., 2012). For studies that report that mobile phone data can be used to infer personalities on the basis of the calling patterns of phone users (de Montjoye et al., 2013b), the rationale for personality inference is problematic, which is why the accuracy of personality inference is undesirable.

Moreover, the geo-location information in mobile phone data is not the exact location of phone calling activities but the location of mobile phone towers wherein the mobile phone positioning network is built (Gao et al., 2013; Zuo and Zhang, 2012). The geo-location accuracy for phone activities depends on the density of these towers and signal strength which vary considerably within a city (Gao et al., 2013). Furthermore, the data can identify only working and residential activities during weekdays and recreation activities during weekends. The problem with activity identification is that researchers generally use an arbitrary pre-defined activity time to differentiate these activities, thus adding uncertainties to the conclusions of studies using the arbitrary parameters. In addition, mobile phone data only record the moving patterns of people, which is only a small part of the daily life of people, but ignore most of the time spent in the office or at home. These activities are much more closely related to the behavior, health, social, and economic activities of urban residents, which

Please cite this article in press as: Liu, J., et al. Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006



a. The real relationship between x and y

b. The erroneous fitting curve with a noised big data at finer granularity

Fig. 2. Fine-grained big data could cause over-fitting when wrong models or analysis methods are applied to pursue higher coefficient of determination *r* squared.

are the top concern of the public instead of the human moving patterns. Mobile phone data, however, cannot perform this task because it only contains incomplete and less important information.

Smart card data have similar information incompleteness problems with mobile phone data. Several researchers have adopted a winding approach to obtain the socio-economic data of smart card holders. They combined traditional resident travel survey and land use data to compute the residential addresses of these smart card holders, and then infer their income levels based on the residential addresses (Long and Thill, 2015). These data mining procedures can attain certain accuracy, but introduce huge errors because of many uncertainties involved in the study logic.

A special problem of information noise is associated with the big data analysis. This problem probably occurs, but we have no solid evidences yet. As previously discussed in Section 2.1, big data generally have fine granularity, which provides fine detailed data for research. However, although big data contain information in a finer grained and detailed manner, they also record random variations, fluctuations, and even noise during the measurement. When applying traditional methods such as the machine learning algorithm to analyze big data, researchers can probably run into the phenomenon of over-fitting, where the machine learning algorithm learns from the noise embedded in the fine-grained big data and predicts based on the noised information. Take a simple data fitting task for example, assume that the true relationship between variables *x* and *y* is a linear relationship (Fig. 2a). However, when big data provide fine and more detailed data points as shown in Fig. 2b, the machine learning algorithm will likely come out with a multiple polynomial fitting curve with a higher *r* squared rather than the true linear curve.

3.3. Representativeness problems of big data

The representativeness problem is another consequence of the repurposed big data discussed in Section 3.1. Commercial big data providers generally do not adopt a scientific sampling method when collecting data because of their nature of pursuing commercial profits, which limits the population represented in these data to only a small group of people with low significance and implication. Studies based on these big data that ignore the representativeness problem have probably drawn conclusions that mismatch the claimed population in the studies.

Many of the current big data research that uses social media data are built on the following assumption: social media data have a particularly large number of users, which indicates that the data has a large sample coverage, so the data can represent the entire population or possible population sampling bias can be dismissed (Mayer-Schönberger and Cukier, 2013). However, this scenario is false. The large size and volume of big data do not necessarily mean that the data is random and representative (Boyd and Crawford, 2012), and increasing the quantity does not increase the quality (Cheshire and Batty, 2012). For example, Sina Weibo, the Chinese-localized Twitter, reported that it had 61.4 million active users per day in the 4th quarter of 2013, accounting for only 9.94% of all Internet users and 4.51% of the total population, according to the statistics report on Internet users released by the China Internet Network Information Center (CINIC) (China Internet Network Information Center, 2014b). A survey conducted by CINIC in early 2014 shows that nearly 70% of social media users are under 30 years old (China Internet Network Information Center, 2014a). The population that uses social media is only a small population that is mainly comprised of young people, which is far from being representative of the entire population or even the Internet user population. The representativeness problem of social media data is not only in the age structure, but also in regional divisions. More than one quarter of microblogging users are located in well-developed regions, including Guangdong, Beijing, and Shanghai where the Internet users of these three places only account for 9% of the total population of China's Internet users (Fu and Chau, 2013).

This condition is occurring among social media services all over the world. A 2012 survey that interviewed 1802 American Internet users shows that only 16% of Internet users have a Twitter account, and the majority of Twitter users are African-American, urban residents, and young people between 18 and 29 years old (Duggan and Brenner, 2013). Instagram, which is a popular online photo sharing social media platform, is only used by 13% of the survey participants, and the survey shows that it is only popular among women from Latin America (Duggan and Brenner, 2013).

Mobile phone data also have representativeness problems. Most mobile phone data used in big data studies are from mobile phone users who signed service contracts with a mobile phone operator (Krings et al., 2009). However, many phone users have not signed contracts with a mobile phone operator, whereas others have signed a contract with other mobile phone operators in a city. The phone activities of these users have not been recorded in these data.

Another assumption of these studies based on mobile phone records and social media is that one account or phone number represents one person. However, the truth is that "accounts" and users are clearly never equivalent. Some people share one mobile phone, whereas others have multiple phones for different purposes. For instance, each person in Shanghai reportedly has 1.32 phones on average (Niu et al., 2015). The same case applies in social media accounts because these user-generated contents are not solely

Please cite this article in press as: Liu, J., et al. Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006

produced by humans but by a complex and more-than-human assemblage (Crampton et al., 2013).

3.4. Consistency and reliability problems of big data

Data reliability depends on two aspects. First, the data, their derived measures, and indicators can genuinely represent the facts and information on research subjects without being influenced by the data collector's behavior. Second, the data, their derived measures, and indicators are consistent and stable, regardless of how other unrelated factors change.

However, some big data faile in both aspects of data reliability. Take the social media data for example, studies that use social media data often ignore the fact that the operating company that provides these social media data is a significant confounding variable. The actions and behavior of these companies distort social media data in many ways. For example, when studying the interpersonal communication links between cities, researchers are likely to overlook the fact that the social media services itself alters the true inter-personal networks by recommending friends (Fig. 3) on the basis of the user's place of birth, gender, education background, and other attributes. This personalization feature of "friend recommendation" in social media platforms is similar to the targeted advertising of Google and Baidu, which distract users' attention and distort the original intention of online surfing. This hidden influence exerted by these social media companies undermine the credibility of the conclusion of inter-personal communication or inter-city cyberspace analyses such as (Zhen et al., 2012).

In short, the information in social media big data does not only contain inter-personal or inter-city communication information but also contain the interfering influences of the operating companies behind such social media big data. Using social media big data to analyze the users' behavior and activities without eliminating the influences exerted by the operating companies is problematic.

Search engine data are another example of the distortion of users' behavior and activities. Search engines such as Google generally provide users with a handy function called autocomplete to predict the rest of a word being typed by a user as shown in Fig. 4. This feature is convenient for search engine users, but it distorts users' behavior (Ruths and Pfeffer, 2014). Users can be distracted by searching other words or word combinations instead.

The collection and submission procedures of big data are unreliable. For example, the geo-location of the geo-tagged tweets is probably inaccurate and wrong because Twitter users can post geo-tagged tweets literally at "any place" on Earth. Fig. 5 shows a Twitter user who posted a geo-tagged tweet in Hong Kong, but the same user posted another geo-tagged tweet in New York. How is it possible for a person to jump from Hong Kong to New York in only one minute? Free and without-verification geo-tagged tweet postings are apparently problematic. The collection and submission method of unverified data largely decreases the credibility of studies based on these geo-tagged big data. Several studies that adopted the Twitter user-specified location in users' profile encounter the same verification problem because users can provide a fuzzy location such as "Middle Earth" (Crampton et al., 2013; Graham et al., 2014).

The unreliability issues of big data are also manifested in the instability and inconsistency of big data. The population represented in social media data varies with time, and is never stable and consistent. The number of social media valid users would always change because user tastes change and other social networking platforms emerge. This variation directly changes the demographic structure of its user population. Research based on these social media data can be true now, but can become false a few months later (Ruths and Pfeffer, 2014).

OpenStreetMap, a well-known volunteer geographic information, is an excellent free map data for many researchers (Liu and Long, 2015). However, one problem of OpenStreetMap is that the accuracy and data quality are worrisome. Given that most of the data in OpenStreetMap are provided by different amateur users who have never received professional scientific training on mapping, no uniform and standard data collection methods exist, thus making the data accuracy, quality, and completeness vary considerably across a country or city. The OpenStreetMap data can only be regarded as a "generalized dataset" (Haklay, 2010). In addition, OpenStreetMap appears to "bully the poor and flatter the rich" because it provides detailed map data for wealthy regions, but only rare and incomplete data for poor areas (Haklay, 2010). A reasonable explanation is that the volunteering data collectors of OpenStreetMap tend to collect the data in affluent areas rather than the less affluent areas. The inconsistent and unjust data abundance undermines the reliability of this volunteering geographical big data.

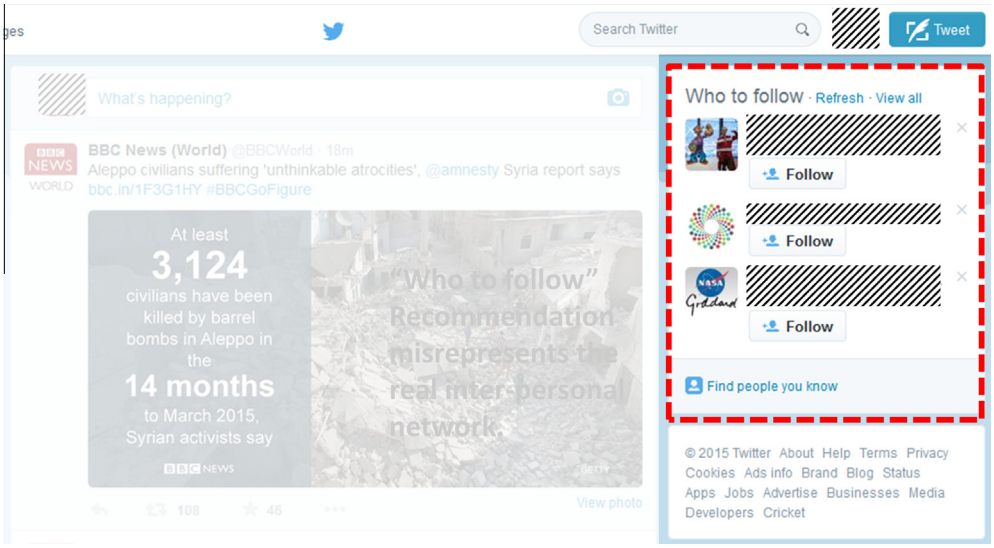


Fig. 3. Social media use friend recommendations and other methods to affect the real social network.

Please cite this article in press as: Liu, J., et al. Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006

Fig. 4. The autocomplete function of Google search engine distorts the user's original search topic.

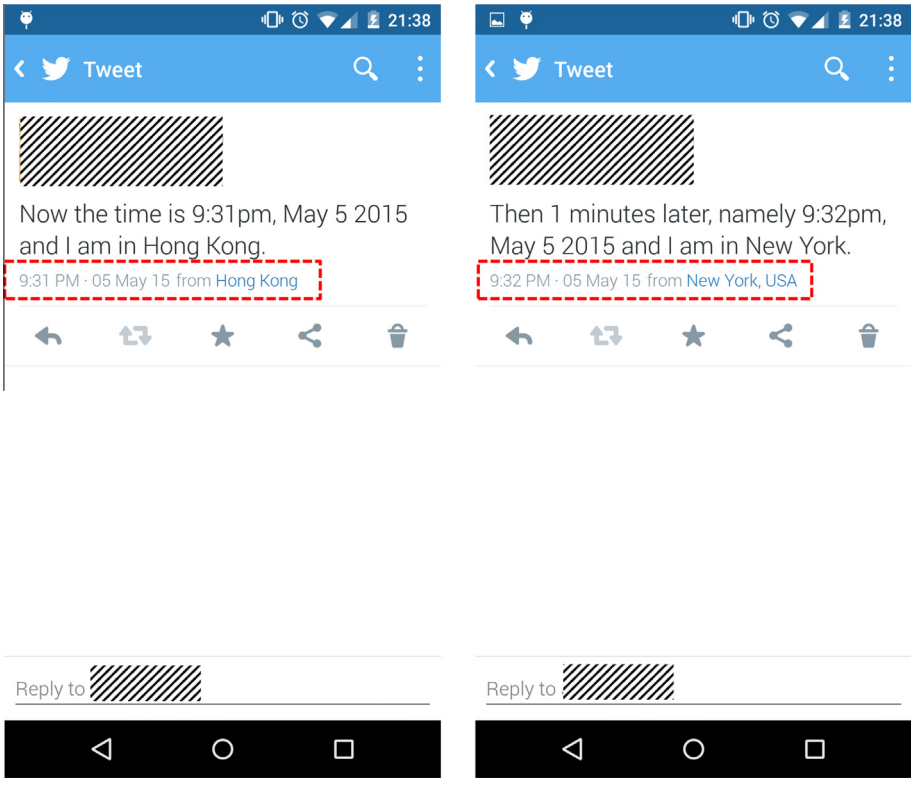


Fig. 5. Geo-tagged tweets can be sent without verification on the geo-location.

3.5. Ethical issues of big data

We once had a discussion with Professor John Logan of Brown University, the author of "Urban Fortunes: The Political Economy of Place." When Professor Logan learned that we used Sina Weibo microblog to collect public activity data, he asked us, "is it ethical?" We replied, "Since the users make public the information, the information is open. So data mining from the information is certainly allowed."

However, the truth is that "just because it is accessible does not make it ethical" and ignoring the ethical evaluation of research is problematic (Boyd and Crawford, 2012). The practice of collecting public data without seeking appropriate ethical approval still has some risks. If the activity pattern of an individual user is found during the data mining process, it can infringe on the user's privacy and exceed the "minimal risk" in research ethics. The so-called minimal risk refers to the pain or discomfort that people experience in studies that should not be more severe than what they experience in daily life (Bacon-Shone, 2014). What we did not consider and acknowledge is the fact that social media data is in public is different from the fact that the permission to use the data is granted by all involved users (Boyd and Marwick, 2011).

Several scholars identified nearly 10,000 people engaged in urban planning and related careers by searching the Sina

Please cite this article in press as: Liu, J., et al. Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006

microblog and analyzed their personal connections (Mao and Long, 2013). Such academic exploration seems fine, but these findings can violate the privacy of the planning practitioners under investigation. For studies that look into the research topics involved with user privacies, does the researcher collect sensitive data regarding social network users' political ideas or actions? Is the research data stored properly? Is there any risk of data leakage? Are the user IDs and names replaced with pseudo codes? Have measures been taken to minimize the minimum risk emphasized in the international research ethics?

Other big data-based studies also reveal the possible violation of big data against user privacy. For example, a study on the communication data of 1.5 million anonymous mobile phone users in a Western European country shows that four spatial–temporal position records can sufficiently confirm the identity of 95% of people; researchers also found that after diluting the time and space of mobile phone data sets, user privacy protection is unimproved (de Montjoye et al., 2013a). The key information in our identification documents is also unsafe, such as social security numbers. Researchers show that individuals' social security numbers can be inferred by combing publicly available data, including profiles in social networking sites (Acquisti and Gross, 2009).

These studies point out the underlying difficult ethical problems in big data. However, these problems are beyond the control of current ethical control mechanisms because we neither know what big data type is liable for violating user privacy nor do we understand what measure can be taken to cope with it. Our research ethics committee is also unprepared for the big data problem (Boyd and Crawford, 2012).

4. Summary and coping strategies

In the previous section, we briefly introduce the 'Big Errors' in data quality and usage problems of big data: inauthentic data collection, information incompleteness, unrepresentativeness, inconsistency and unreliability, as well as ethical issues. Obviously, most of these issues are due to the unscientific practice of data collection, data processing, and the lack of data verification. The 'big errors' brought by Big Data appears critical, but big data still has potential. As aforementioned at the beginning of this study, our aim is not to discourage big data studies. Instead, our intention is to increase the awareness of researchers of the potential biases and errors in this field, as well as to unravel the big data puzzle with care. Big data studies, as currently observed, will continue to prosper and produce more interesting studies.

The next question is how spatial information sciences and related disciplines cope with the "big error" brought by big data. This paper argues that big data studies in spatial information sciences and related disciplines should focus on the following aspects. Note here that the following recommendations are preliminary thoughts on how to mitigate the "Big Error" rather than a panacea for all problems.

4.1. Big data errors should be further understood and evaluated, and new reliable data analysis methods should be developed to decrease errors

Obtaining a deep understanding of the issue is essential to develop appropriate methods to solve the problem. This condition is also applicable for the "big error" brought by big data. Specifically, scholars who use big data should have a deep understanding of how the data supplier's behavior affects the quality of big data and biases the results. Besides, evaluating the quality of big data and possible errors, such as positioning accuracy, data integrity, logical consistency, and other data accuracy characteristics, is needed before using big data in data analysis. Second, we should actively develop reliable technical methods that can be widely applied to decrease or eliminate big data errors. When countermeasures to solve these problems are unrealistic, we should carefully determine the scope of study such as the targeted study population or subjects, and explain it in the discussion section.

4.2. We should cooperate with data providers, adopt rigorous research designs such as experimental research design to decrease or eliminate the influence from data providers

The impacts of big data suppliers on the big data quality have been discussed in the previous part. If researchers who know the scientific methods of data collection cooperate with big data suppliers and conduct rigorous research plans together, they are likely to obtain a conclusions with high credibility. This model is feasible because researchers need the data and support from big data suppliers and the suppliers want to learn more about their business from the data. For example, eBay may want to know how it can influence online transactions and logistics if they increase trading commissions for the merchants. There are precedents of scholars in other fields that cooperate with online commercial platforms. For example, Kohavi et al. (2009), who were the experimental platform team members of Microsoft, cooperated with Amazon, Google, and NASA using an experiment design method to study user satisfaction or tolerance on the waiting time of the searching web, certain website designs, or certain services.

4.3. Research on big data should be supplemented by traditional scientific data collection methods to obtain more detailed and representative data

Big data such as mobile phone data and smart card data have incompleteness problems and other big data such as social media microblogs have biased samples. Traditional data can be used to complement such big data. Using a scientific traditional sampling survey, we can collect more detailed information of the target population including as their socio-economic characteristics, and make the collected data more representative. Many scholars are already using this approach, such as combining smart card data with travel diaries (Long et al., 2015) and combining news report and crime incident data with social media data (Crampton et al., 2013).

4.4. Multiple data sources can be used to expand the sample representativeness and enhance the reliability of research findings based on big data

Data from multiple sources result in relatively complex data structures and data processing tasks, but it increases diversity. This condition is particularly important for the problems of representativeness and reliability of big data. The robustness of conclusions derived from one type of big data can be tested with big data provided by another platform to make them more convincing.

4.5. Current research ethics and good practices should be enhanced at the governmental, university, and individual levels all over the world

As providers of public goods, government agencies, universities, research institutions, and individual researchers are obliged to protect the privacy of individuals and the right of the public to know. Ethics review practice is well implemented in developed countries but poorly implemented in developing countries such as China. To the best of our knowledge, clinical research ethics reviews are widely adopted in the medical schools of Chinese universities but there are no compulsory institutional-level measure over non-clinical research that involves human subjects. We believe that

Please cite this article in press as: Liu, J., et al. Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006

the research ethics review is not to hinder the progress of research. Instead, it attempts to protect the public, research institutions, and researchers themselves.

As for big data research, we recommend that access control to big data that are liable to "disclosing" the privacy of the public should be regulated in big data research. Pseudo codes should be used to substitute all identifiable information. However, what kind of big data is vulnerable to privacy disclosure and what corresponding countermeasures can be developed should be intensively studied first.

Acknowledgement

This research was support by the Early Career Scheme from Research Grant Council of Hong Kong (Project No.: 27200414).

References

Acquisti, A., Gross, R., 2009. Predicting social security numbers from public data. Proc. Natl. Acad. Sci. 106, 10975–10980.
Anderson, C., 2008. The end of theory: the data deluge makes the scientific method obsolete. Wired Mag.
Bacon-Shone, J., 2014. Human Research Ethics in HKU. <http://intraweb.hku.hk/local/rss/RCR/human-research-ethics-hku.pdf> (accessed 16.12.15).
Batty, M., 2012. Smart cities, big data. Environ. Plan. B – Plan. Des. 39, 191–193. http://dx.doi.org/10.1068/b3902ed.
Boyd, D., Crawford, K., 2012. Critical questions for big data provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15, 662–679. http://dx.doi.org/10.1080/1369118x.2012.678878.
Boyd, D., Marwick, A.E., 2011. Social privacy in networked publics: Teens' attitudes, practices, and strategies. In: Proceedings of the a Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. Oxford Internet Institute, pp. 1–29.
Cheshire, J., Batty, M., 2012. Visualisation tools for understanding big data. Environ. Plan. B – Plan. Des. 39, 413–415. http://dx.doi.org/10.1068/b3903ed.
China Internet Network Information Center, 2014a. Research Report on Social Media User Behaviors 2014. <https://www.cnnic.cn/hlwfzyj/hlwxzbg/sqbg/201408/P020150401351309648557.pdf> (accessed 16.12.15).
China Internet Network Information Center, 2014b. Statistical Report on Internet Development in China. <http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201403/P020140305346585959798.pdf> (accessed 16.12.15).
Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W., Zook, M., 2013. Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. Cartogr. Geogr. Inf. Sci. 40, 130–139. http://dx.doi.org/10.1080/15230406.2013.777137.
de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013a. Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. 3. http://dx.doi.org/10.1038/srep01376.
de Montjoye, Y.-A., Quoidbach, J., Robic, F., Pentland, A., 2013b. Predicting personality using novel mobile phone-based metrics. In: Greenberg, A., Kennedy, W., Bos, N. (Eds.), Social Computing, Behavioral–Cultural Modeling and Prediction. Springer Berlin Heidelberg, pp. 48–55.
Duggan, M., Brenner, J., 2013. The Demographics of Social Media Users – 2012. <http://www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012/> (accessed 16.12.15).
Einav, L., Levin, J., 2014. Economics in the age of big data. Science 346, 1243089. http://dx.doi.org/10.1126/science.1243089.
Fu, K.-W., Chau, M., 2013. Reality check for the Chinese microblog space: a random sampling approach. PLoS ONE 8, e58356.
Gao, S., Li, L., Li, W., Janowicz, K., Zhang, Y., 2014. Constructing gazetteers from volunteered big geo-data based on Hadoop. Comput. Environ. Urban Syst. http://dx.doi.org/10.1016/j.compenvurbsys.2014.02.004.
Gao, S., Liu, Y., Wang, Y., Ma, X., 2013. Discovering spatial interaction communities from mobile phone data. Trans. GIS 17, 463–481.
Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L., 2008. Understanding individual human mobility patterns. Nature 453, 779–782. http://dx.doi.org/10.1038/nature06958.
Graham, M., Hale, S.A., Gaffney, D., 2014. Where in the world are you? Geolocation and language identification in Twitter. Prof. Geogr. 66, 568–578.
Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environ. Plan. B – Plan. Des. 37, 682–703.
Huang, W., Li, S., Liu, X., Ban, Y., 2015. Predicting human mobility with activity changes. Int. J. Geogr. Inf. Sci. 29, 1–19.
IBM, 2013. The Four V's of Big Data. <http://www.ibmbigdatahub.com/infographic/four-vs-big-data> (accessed 16.12.15).
Khoury, M.J., Ioannidis, J.P.A., 2014. Big data meets public health. Science 346, 1054–1055. http://dx.doi.org/10.1126/science.aaa2709.
Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.M., 2009. Controlled experiments on the web: survey and practical guide. Data Min. Knowl. Disc. 18, 140–181.
Krings, G., Calabrese, F., Ratti, C., Blondel, V.D., 2009. Urban gravity: a model for inter-city telecommunication flows. J. Stat. Mech: Theory Exp. 2009, L07003.
Laney, D., 2001. 3-D Data Management: Controlling Data Volume, Velocity and Variety. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> (accessed 16.12.15).
Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., 2009. Life in the network: the coming age of computational social science. Science 323, 721.
Lazer, D.M., Kennedy, R., King, G., Vespignani, A., 2014. The parable of Google Flu: traps in big data analysis. Science 343, 1203–1205.
Lee, R., Wakamiya, S., Sumiya, K., 2013. Urban area characterization based on crowd behavioral lifelogs over Twitter. Pers. Ubiquit. Comput. 17, 605–620. http://dx.doi.org/10.0007/s00779-012-0510-9.
Liu, X., Long, Y., 2015. Automated identification and characterization of parcels (AICP) with OpenStreetMap and Points of Interest. Environ. Plan. B – Plan. Des., 1–20 http://dx.doi.org/10.1177/0265813515604767.
Liu, Y., Kang, C., Gao, S., Xiao, Y., Tian, Y., 2012. Understanding intra-urban trip patterns from taxi trajectory data. J. Geogr. Syst. 14, 463–483.
Long, Y., Liu, X., Zhou, J., Chai, Y., 2015. Early Birds, Night Owls, and Tireless/Recurring Itinerants: An Exploratory Analysis of Extreme Transit Behaviors in Beijing, China. arXiv:1502.02056 [physics.soc-ph].
Long, Y., Thill, J., 2015. Combining smart card data, household travel survey and land use pattern for identifying housing-jobs relationships in Beijing. Comput. Environ. Urban Syst. 53, 19–35. http://dx.doi.org/10.1016/j.compenvurbsys.2015.02.005.
Loshin, D., 2012. Data Governance and Quality: Data Reuse vs. Data Repurposing. <http://dataqualitybook.com/?p=349> (accessed 06.05.15).
Mao, M., Long, Y., 2013. Ji yu wei bo shu ju de gui hua quan shi bie chu tan [an exploration into the personal connection of urban planning practitioners using Sina Weibo data]. In: Annual National Planning Conference 2013. Urban Planning Society of China, Qingdao, China.
Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution that Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.
McNulty, E., 2014. Understanding Big Data: The Seven V's <http://dataconomy.com/seven-vs-big-data/> (accessed 15.12.15).
Niu, X., Ding, L., Song, X., 2015. Understanding urban spatial structure of shanghai central city based on mobile phone data. China City Planning Review 3, 004.
O'Leary, D.E., 2013. Artificial intelligence and big data. IEEE Intell. Syst. 28, 96–99.
Pei, T., Sobolevsky, S., Ratti, C., Amini, A., Zhou, C., 2014. Uncovering the directional heterogeneity of an aggregated mobile phone network. Trans. GIS 18, 126–142.
Pigliucci, M., 2009. The end of theory in science? EMBO Rep. 10, 534–534.
Pijanowski, B.C., Tayyebi, A., Doucette, J., Pekin, B.K., Braun, D., Plourde, J., 2014. A big data urban growth simulation at a national scale: configuring the GIS and neural network based land transformation model to run in a high performance computing (HPC) environment. Environ. Modell. Softw. 51, 250–268. http://dx.doi.org/10.1016/j.envsoft.2013.09.015.
Reades, J., Calabrese, F., Ratti, C., 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. Environ. Plan. B – Plan. Des. 36, 824–836.
Richardson, D.B., Volkow, N.D., Kwan, M.-P., Kaplan, R.M., Goodchild, M.F., Croyle, R.T., 2013. Spatial turn in health research. Science 339, 1390.
Robinson, W.S., 2009. Ecological correlations and the behavior of individuals. Int. J. Epidemiol. 38, 337–341.
Roth, C., Kang, S.M., Batty, M., Barthélemy, M., 2011. Structure of urban movements: polycentric activity and entangled hierarchical flows. PLoS ONE 6, e15923.
Ruths, D., Pfeffer, J., 2014. Social media for large studies of behavior. Science 346, 1063–1064. http://dx.doi.org/10.1126/science.346.6213.1063.
Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010. Limits of predictability in human mobility. Science 327, 1018–1021.
Soto, V., Frías-Martínez, E., 2011. Automated land use identification using cell-phone records. In: Proceedings of the 3rd ACM International Workshop on MobiArch. ACM, pp. 17–22.
Toole, J.L., Ulm, M., González, M.C., Bauer, D., 2012. Inferring land use from mobile phone activity. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing. ACM, pp. 1–8.
Webster, C., 2014. Dean's Roundup (Friday, 31 October, 2014). <http://fac.arch.hku.hk/wp-content/uploads/2014/10/DeansRoundup-20141031.pdf> (accessed 06.05.15).
Wu, H.Y., Zhang, T., Gong, J.Y., 2014. Geocomputation for geospatial big data. Trans. GIS 18, 1–2. http://dx.doi.org/10.1111/tgis.12131.
Yuan, J., Zheng, Y., Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the ACM KDD. ACM, pp. 186–194.
Zhen, F., Wang, B., Chen, Y., 2012. China's city network characteristics based on social network space: an empirical analysis of sina micro-blog. Acta Geogr. Sin. 67, 1031–1043.
Zheng, Y., Liu, F., Hsieh, H.-P., 2013. U-Air: when urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1436–1444.
Zuo, X., Zhang, Y., 2012. Detection and analysis of urban area hotspots based on cell phone traffic. J. Comput. 7, 1753–1760.

Please cite this article in press as: Liu, J., et al. Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.11.006