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/© 2015 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

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.

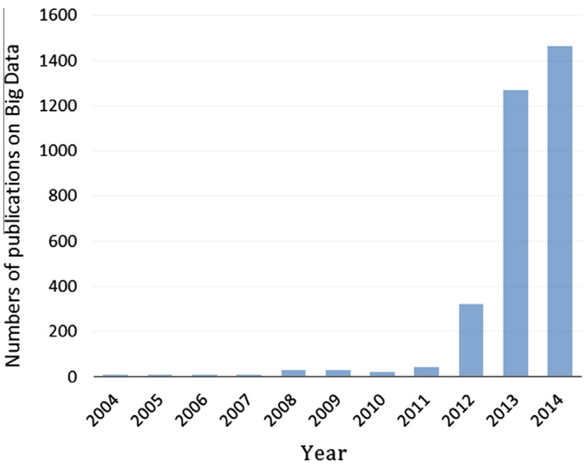

Fig. 1. Number of published studies on “big data” based on literature search of the phrase “big data” in the topic field in the database of web of knowledge from the year 1956 to 2014.

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