Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Review Article

Geospatial big data handling theory and methods: A review and research challenges

Songnian Li[a,*], Suzana Dragicevic[b], Francesc Antón Castro[c], Monika Sester[d], Stephan Winter[e], Arzu Coltekin[f], Christopher Pettit[g], Bin Jiang[h], James Haworth[i], Alfred Stein[j], Tao Cheng[i]

[a] Ryerson University, Toronto, Canada
[b] Simon Fraser University, Burnaby, Canada
[c] Technical University of Denmark, Lyngby, Denmark
[d] Leibniz University Hannover, Germany
[e] University of Melbourne, Australia
[f] University of Zurich, Switzerland
[g] University of New South Wales, Australia
[h] University of Gävle, Sweden
[i] University College London, UK
[j] University of Twente, The Netherlands

ARTICLE INFO

Article history:
Received 31 May 2015
Received in revised form 13 October 2015
Accepted 30 October 2015
Available online xxxx

Keywords:
Big data
Geospatial
Data handling
Analytics
Spatial modeling
Review

ABSTRACT

Big data has now become a strong focus of global interest that is increasingly attracting the attention of academia, industry, government and other organizations. Big data can be situated in the disciplinary area of traditional geospatial data handling theory and methods. The increasing volume and varying format of collected geospatial big data presents challenges in storing, managing, processing, analyzing, visualizing and verifying the quality of data. This has implications for the quality of decisions made with big data. Consequently, this position paper of the International Society for Photogrammetry and Remote Sensing (ISPRS) Technical Commission II (TC II) revisits the existing geospatial data handling methods and theories to determine if they are still capable of handling emerging geospatial big data. Further, the paper synthesises problems, major issues and challenges with current developments as well as recommending what needs to be developed further in the near future.

1. Introduction

Over the last decade, big data has become a strong focus of global interest, increasingly attracting the attention of academia, industry, government and other organizations. The term “big data” first appeared in the scientific communities in the mid-1990s, gradually became popular around 2008 and started to be recognized in 2010. Today, big data is a buzzword everywhere on the Internet, in the trade and scientific publications and during all types of conferences. Big data has been suggested as a predominant source of innovation, competition and productivity (Manyika et al., 2011), and has caused a paradigm shift to data-driven research (Kitchin, 2014). The rapid growing flood of big data, originating from the many different types of sensors, messaging systems and social networks in addition to more traditional measurement and observation systems, have already invaded many aspects of our everyday existence. On the one hand, big data, including geospatial big data, has great potential to benefit many societal applications such as climate change, disease surveillance, disaster response, monitoring critical infrastructures and transportation. On the other hand, big data’s benefits to society are usually limited by issues such as data privacy, confidentiality and security.

Big data is still not a clearly defined term and it has been defined differently from technological, industrial, research or academic perspectives (Chen et al., 2014). In general, it is considered as structured and unstructured datasets with massive data volumes that cannot be easily captured, stored, manipulated, analyzed, managed and presented by traditional hardware, software and database technologies. Along with its definitions, big data is

* Corresponding author.
E-mail addresses: snli@ryerson.ca (S. Li), suzanad@sfu.ca (S. Dragicevic), fa@space.dtu.dk (F.A. Castro), monika.sester@ikg.uni-hannover.de (M. Sester), winter@unimelb.edu.au (S. Winter), arzu@geo.uzh.ch (A. Coltekin), c.pettit@unsw.edu.au (C. Pettit), bin.jiang@hig.se (B. Jiang), j.haworth@ucl.ac.uk (J. Haworth), a.stein@utwente.nl (A. Stein), tao.cheng@ucl.ac.uk (T. Cheng).

http://dx.doi.org/10.1016/j.isprsjprs.2015.10.012
0924-2716/© 2015 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

Please cite this article in press as: Li, S., et al. Geospatial big data handling theory and methods: A review and research challenges. ISPRS J. Photogram. Remote Sensing (2015), http://dx.doi.org/10.1016/j.isprsjprs.2015.10.012

often described by its unique characteristics. In discussing application delivery strategies under increasing data volumes, Laney (2001) first proposed three dimensions that characterize the challenges and opportunities of increasing large data volumes: *Volume*, *Velocity* and *Variety* (3Vs). While *3Vs* have been continuously used to describe big data, the additional dimension of *Veracity* has been added to describe data integrity and quality. Further *Vs* have also been suggested such as variability, validity, volatility, visibility, value, and visualization. However, these are met critically as they do not necessarily express qualities of magnitude. While it is true these further *Vs* do not directly contribute to understanding the "big" in big data, they do touch on important concepts related to the entire pipeline of big data collection, processing and presentation. Suthaharan (2014) even argued that 3Vs cannot support early detection of big data characteristics for its classification and proposed 3Cs: *cardinality*, *continuity*, and *complexity*. It is apparent that defining big data and its characteristics will be an ongoing endeavour, but it nevertheless will not have negative impact on big data handling and processing.

According to the arguable phrase "80% of data is geographic" (see discussions in Morais (2012)), much of the data in the world can be geo-referenced, which indicates the importance of geospatial big data handling. Geospatial data describe objects and things with relation to geographic space, often with location coordinates in a spatial referencing system. Geospatial data are usually collected using ground surveying, photogrammetry and remote sensing, and more recently through laser scanning, mobile mapping, geo-located sensors, geo-tagged web contents, volunteered geographic information (VGI), global navigation satellite system (GNSS) tracking and so on. Adopting the widely accepted characterization method, geospatial data can exhibit at least one of the *3Vs* (Evans et al., 2014), but the other *Vs* mentioned above are also relevant. As such, geospatial big data can be characterized by the following, with the first four being more fundamental and important:

- *Volume:* Petabyte archives for remotely sensed imagery data, ever increasing volume of real-time sensor observations and location-based social media data, vast amount of VGI data, etc., as well as continuous increase of these data, raise not only data storage issues but also a massive analysis issue (Dasgupta, 2013).
- *Variety:* map data, imagery data, geotagged text data, structured and unstructured data, raster and vector data, all these different types of data – many with complex structures – calls for more efficient models, structures, indexes and data management strategies and technologies, e.g., use of NoSQL.
- *Velocity:* imagery data with frequent revisits at high resolution, continuous streaming of sensor observations, Internet of Things (IoT), real-time GNSS trajectory and social media data all require matching the speed of data generation and the speed of data processing to meet demand (Dasgupta, 2013).
- *Veracity:* much of geospatial big data are from unverified sources with low or unknown accuracy, level of accuracy varies depending on data sources, raising issues on quality assessment of source data and how to "statistically" improve the quality of analysis results.
- *Visualization:* provides valuable procedures to impose human thinking into big data analysis. Visualizations help analysts identifying patterns (such as outliers and clusters), leading to new hypotheses as well as efficient ways to partition the data for further computational analysis. Visualizations also help end users to better grasp and communicate dominant patterns and relationships that emerge from the big data analysis.
- *Visibility:* the emergence of cloud computing and cloud storage has made it possible to now efficiently access and process geospatial big data in ways that were not previously possible. Cloud technology is still evolving and once issues such as data provenance – historical metadata – are resolved, big data and the cloud would be mutually dependent and reinforcing technologies.

The increasing volume and varying format of collected geospatial big data pose additional challenges in storing, managing, processing, analyzing, visualizing and verifying the quality of data. Shekhar et al. (2012, p. 1) states that "the size, variety and update rate of datasets exceed the capacity of commonly used spatial computing and spatial database technologies to learn, manage, and process the data with reasonable effort". Big data tends to hold people to expect more and larger hypotheses that grow faster than the statistical strength of data and capacity of data analysis (Gomes, 2014). Verifying the quality of geospatial big data and data products delivered to end users is noted as one of the big challenges and becomes even more challenging in the quality control of the delivered data products (see 2012 ISPRS Resolution, www.isprs.org/documents/resolutions.aspx). On the other hand, fitness of uses or purposes appears more valid or should be advocated (Mayer-Schönberger and Cukier, 2013) in the context of big data.

The objectives of this paper are to (1) revisit the existing geospatial data handling methods and theories to determine if they are still capable of handling emerging geospatial big data; (2) examine current, state-of-the-art methodological, theoretical, and technical developments in modeling, processing, analyzing and visualizing geospatial big data; (3) synthesize problems, major issues and challenges in current developments; and (4) recommend what needs to be developed in the near future. Sections 2–6 addresses objectives 1 and 2 of the 5 important areas related to geospatial big data handling methods and theories, which are the focus of various Working Groups (WG) of ISPRS TC II. Related image analysis and processing topics, such as dimensionality reduction; image compression; compressive sensing in big data analytics; content-based image retrieval; and image endmember extraction, are not covered in this paper. Section 7 presents open issues and future research directions of the three focus areas of TC II. Section 8 gives a summary and conclusions to the paper.

2. Collection of geospatial big data

In recent years, along with the availability of new sensors, new ways of collecting geospatial data have emerged, leading to completely new data sources and data types of geographical nature. Data acquired by the public, so-called Volunteered Geographic Information (VGI), and data from geo-sensor networks have led to an increased availability of spatial information. Whereas until recently, authoritative datasets were dominating in topographic domain, these new data types extend and enrich geographic data in terms of thematic variation and by the fact that it is more user-centric. The latter is especially true for VGI collected by social media (Sester et al., 2014).

Geospatial data collection is shifting from a data sparse to a data rich paradigm. Whereas some years back geospatial data capture was based on technically demanding, accurate, expensive and complicated devices, where the measurement process was itself sometimes an art, we are now facing a situation where geospatial data acquisition is a commodity implemented in everyday devices such as smartphones used by many people. These devices are capable of acquiring environmental geospatial information at an unprecedented level with respect to greatly improved geometric