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

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/

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

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

accuracy, temporal resolution and thematic granularity. They are small, easy to handle, and able to acquire data even unconsciously.

This data capture paradigm is similar to the situation in topographic data collection for digital terrain models by capturing significant topographic points with morphological characteristics on the one hand ("qualified" points, i.e., points with semantics) – as opposed to the collection of point clouds using LiDAR sensors or stereo matching, leading to masses of "unqualified" points (Ackermann, 1994). The first approach requires manual selection and measurement and guarantees that the topographic reality can be interpolated from the sparse measurements. The second approach assumes that the topographic reality is captured by the dense measurements and can be reconstructed from them – thus the object formation and identification are shifted to the analysis process.

In general, one can distinguish the following sensor configurations: (1) objects equipped with sensors moving through space and capturing their own trajectories and the local environments: humans and moving devices such as cars; and (2) static sensors constantly observing the (changing) environment. Today's data acquired by these new sensors and new stakeholders can be characterized as follows:

- data streams,
- arbitrary high density,
- "close sensing" (Duckham, 2013), i.e., the ability to measure many different dimensions of objects characteristics, e.g., optical, acoustic, and mechanical features, and
- different degrees of positional accuracy and reference, ranging from highly precise coordinates via relative positions to information where there is no geometric reference or is only implicitly located by location names.

There are many examples of data collections that may lead to geospatial big data sets. For example, from a "social" perspective, over the last decade we have seen (through the rise of the so-called "smart city" concept) the instrumentation of cities which are now providing vast amount of real-time data through the likes of smart card ticketing systems, vehicle tracking devices, CCTV, toll systems, induction loops and other sensors. With the rise of social media we are also seeing vast amounts of data (e.g., Twitter feeds), which can be geotagged and used to assist in disaster management and emergency relief. From an "environmental science" perspective, there are huge remotely sensed imagery repositories such as NASA's Landsat repositories which provide petabytes of geospatial data (Riebeek, 2015). Capturing the urban environment is possible using a variety of sensors. Cars (e.g., connected vehicles) are equipped with many sensors to aid the driver and enhance safety and comfort. These sensors also capture the immediate environment of the car using front cameras, backwards cameras, ultrasonic (for parking assistance), GPS, radar, rain-sensing wipers (Fitzner et al., 2013). The information is stored on local systems and can be transmitted to the available infrastructure or other vehicles.

Another important example concerns the quality of data on health. Such data are routinely collected and stored, e.g., with doctors or health centres. In particular in public health centres, however, the coordinates in space and time may lack quality. The first reason is that health aspects are not always related to the location where the person lives. The second reason is that the moment when he or she is visiting the health facility may not correspond with the time of incidence. As an important reason we found that people avoid stigmatic investigations, e.g., related to AIDS or other sexually transmitted diseases, thus giving a bias in routinely collected datasets (Kandwal et al., 2010).

People, considered as "sensors", can also help capture traffic or mobility related VGI style information. VGI acquisition can be distinguished into participatory and opportunistic. Participatory data acquisition is conducted in a conscious process by a user, who explicitly selects objects and their features and contributes this information (an example is the OpenStreetMap, OSM). Opportunistic data capture occurs unconsciously, mostly with no specific purpose – or even a completely different purpose. A prominent example is the exploitation of mobile phone data to determine traffic information such as traffic jams. The capture of the spatio-temporal phenomenon "traffic jam" is just a by-product of many users having switched their phones on when driving their cars. With the recent emergence of smart cards, transport ticketing systems like the London Oyster card are capturing the movement of millions of travellers which use the London Tube and railway system daily. A new data point is created every time they register the location via swiping on or off a mode of public transport.

VGI has proven to be an essential data source, when ad-hoc information is needed as in a disaster context (Goodchild and Glennon, 2010). In those situations, it is important to get any information – even if it is not very accurate. Thus, social media and services are considered as new information sources for example for early response and crisis management (Fuchs et al., 2013). Fuchs et al. (2013) evaluated Twitter streams to detect large scale flooding events in Germany. In a period of 8 months, approximately 6 million tweets had been recorded. If the analysis concentrates only on the frequency, it was not possible to identify the events; however, the inclusion of specific keywords, together with spatio-temporal clustering was able to detect some of the events. A similar approach is reported by Dittrich and Lucas (2013). Huang et al. (2015) used millions of location-based tweets to predict human movements. Other examples for the successful use of crowdsourcing data collection approach in the context of disasters, are the Haiti earthquake (http://www.ushahidi.com/), the Queensland flood (McDougall, 2011), as well as flood risk assessment (Poser et al., 2009). Invitation active participation of users in the context of mapping has been reported by Frommberger et al. (2013).

It is worth noting that in the realm of geoscientific data, there is a wealth of new sensors and data sources, which lead to large collections of diverse and "dirty" data (inaccurate, incomplete or erroneous data), which only gain relevance by careful integration and fusion with complementary data (Van Zyl et al., 2009).

3. Quality assessment

Geospatial data, being abstractions and observations of a continuous reality (Frank, 2001), is by nature uncertain, ideally time-stamped and often incomplete. Hence, geospatial big data, with its defining characteristics of being large (voluminous), heterogeneous (variety), real-time processed (velocity), inconsistent (variability), and thus also of variable quality (veracity), must suffer even more from uncertainty, asynchronicity, and incompleteness. However, while certain effects on data quality are emphasized for geospatial big data, the phenomena to be described are still the same. Thus, the known methods and theories of quality assessment are still applicable.

In geographic information science and technology, standardized methods have been developed in order to assess, describe and propagate quality characteristics both quantitatively and qualitatively. Frameworks exist to describe data quality from a producer's perspective, which then, in the hands of consumers, have to be translated into fitness for purpose. These frameworks went into international standards, such as ISO 8402 (which is generically about quality management: "Data quality is the

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

totality of characteristics of a product that bear on its ability to satisfy stated and implied needs") or ISO 19157 (which is specifically about the quality of geographic information: "Establishes the principles for describing the quality of geographic data. It defines components for describing data quality, specifies components and content structure of a register for data quality measures, describes general procedures for evaluating the quality of geographic data, and establishes principles for reporting data quality"). The frameworks typically define quantitative measures of data quality, such as spatial, temporal and thematic accuracy, spatial, temporal and thematic resolution, consistency, and completeness (Veregin, 2005), and in addition qualitative characteristics of data quality, such as purpose, usage, or lineage.

In the end, this approach to data quality assessment is descriptive about the data capturing process, stored in separate metadata. But these data are used in decision making processes, and thus, the quantities and qualities of the above frameworks require further analysis for propagation. In a spatial statistical context work has been done by Van de Vlag et al. (2005) and Van de Vlag and Stein (2006) on natural objects and Kohli et al. (2012) on slums. Frank (2007, p. 417), studying the ontology of imperfect knowledge, stated: "How do the imperfections in the data affect the decision?", but acknowledges that the decision making process is a black box of unknown complexity. While still some quality descriptors, especially the quantitative ones, lend themselves to functional error propagation, others – especially the qualitative ones – are still left to human judgment. With the ad-hoc combination of data streams in geospatial big data collection and near-real-time analytics, this traditional approach falls short both in collecting, aggregating or propagating metadata as well as in human judgment of metadata. These challenges can be illustrated for example in the context of connected urban transport introduced above:

- The sheer volume of geospatial big data in transport arises from the large number of agents (vehicles, people, and goods) on their way at any time. Furthermore, in order to be able to predict transport demand or traffic, not only are real-time data required but also historic data. If the bandwidth of communication channels forms bottlenecks, even with lossless compression, either the sampling rate or the sampling size can be reduced (loss of details), or the computation can be decentralized (Duckham, 2013), in which case a central instance would collect only aggregated data. Each of these solutions has an impact on the quality assessment of the collected data.
- The large variety of geospatial big data in transport (such as GNSS, inertial sensors, compass, wheel sensors, radar, laser scanning, number plate recognition, induction loops, electronic toll or ticketing, parking sensors, social media comments, citizen reports, to name only a few) to be combined in big data analytics challenges error propagation models, which are based on a functional relationship.
- The velocity of data, more often than not requiring real-time analysis for event management or near-future prediction, does not allow for setting up proper error propagation, since disturbances in the sensor or channel can only be detected in hindsight or from prediction models (Umamaheshwaran et al., 2007).
- The variability of geospatial big data in transport, indicating inconsistency, stems from both the variety between different data channels as well as the unreliability of any of these sources. For example, volunteered geographic information can be inconsistent if the "citizens as sensors" (Goodchild, 2007) or "citizens as databases" (Richter and Winter, 2011) disagree on their observations, are just not sampling a particular phenomenon of interest (e.g., if a tree falls on a street and nobody is there to notice), or are limited by lack of communication channels (e.g., a sensor is moving out of reach of WiFi/cell phone coverage). In another example, satellite positioning has some well-known quality descriptors, but in urban canyons, these measures vary with the location in the environment rather than the sensor, the transmission channel or the time (Kealy et al., 2014). This dependency on location is non-linear and difficult to predict.
- The veracity of geospatial big data in transport indicates already that data is to be combined of very different quality. This extends also to irregular sampling rates (both spatial and temporal), entry errors, redundancy, corruption, lack of synchronization, or a variety of collection purposes (taxonomies, semantics), to name a few (Stein et al., 1998).
- Variability and veracity are closely tied to vulnerability, and vulnerability is perhaps the only aspect that is in contrast to classic institutional databases, which are behind firewalls or even completely disconnected (in safety critical applications). Since both big data collection and analytics require connectivity, concepts are also required to deal with malicious contributions, attacks and theft, and privacy. The latter is particularly true for geospatial data collected for tracking movements, and a measure of quality would cater for protection of privacy while still guaranteeing some level of quality of service (Anthony et al., 2007; Duckham and Kulik, 2006).

Within geospatial big data in transport, a prominent example is research that observes the quality of OpenStreetMap, which has become an open source for navigation services at a global scale. This research focuses mostly on the completeness of OpenStreetMap (Haklay, 2010; Mondzech and Sester, 2011; Neis et al., 2012; Zielstra and Zipf, 2010).

Typically, the value of geospatial big data (analytics) is seen in information for decision support. Analytical methods such as data mining and machine learning enable only inductive reasoning on big data, i.e., detection of global correlations, or predictions based on these correlations. In transport, an example is the early discovery of traffic accidents. In these applications the traditional quality (from a provider or consumer perspective) has been replaced by correlation coefficients (Miller and Goodchild, 2014), well knowing that correlation is not necessarily about causes or truth. Thus, validity or trust is traded for the velocity of information production.

4. Data modeling and structuring

All the spatial data models, including the spaghetti vector data model, the network data model, the topological data model and the regular (raster) and irregular (Voronoi, k-d-tree, Binary Space Partitioning Tree, paving group, crystallographic group) tessellations based data models, can be used for handling geospatial big data. Nevertheless, there are some models that are more suitable for handling very large data sets and others that are less suitable for geospatial big data. As the network and topological data models need to store the connectivity (for the network and topological spatial data model) and the adjacency (for the topological spatial data model), they are not well suited for handling geospatial big data streams unless a very efficient spatial data indexing is making the update of the connectivity and/or the topology possible in real time. Since geospatial big data environments might have to relax accuracy constraints to satisfy the real-time constraint, irregular and especially regular tessellations are ideally suited to handle soft errors – geometric uncertainties in data that do not cause erroneous topology. Finally, as regular tessellations can be stored in matrices, they are subject to very parallelized vectorization

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

algorithms. Nevertheless, the implicit topology of the regular tessellation (raster) spatial data model might not be desired topology in some applications that require small inaccuracies (e.g., collision detection for driver-less cars). In such cases, the only possible way to satisfy real time constraints is to use a spatial data indexing method that can maintain its performances with a big data stream being updated in real time. Current spatial data indexing methods cannot handle geospatial big data streams, because their efficiency gets lower as new spatial data streams go over the capacity of extension of the spatial data index (e.g., all the available locations for a hash function value get spent and the hash function needs to get updated or a tree must be rebalanced after a series of additions of data from the stream).

Spatial statistics is well suited to handle big data. It offers capabilities to summarize the data, and express measures of variation and uncertainty. The main concern, however, is that many of the processes and procedures are developed for smaller datasets. In particular, much of spatial statistical analysis is either done on datasets that are collected at a pointwise scale (such as field data, meteorological data, or administrative data) and of a relatively small content, or is focusing on the relatively large image data sets which have a very specific nature. Spatial statistics depends upon the notion of spatial (and spatio-temporal) dependence, and such dependence in turn depends upon the notion of distance between points. For n observations, including their coordinates in space or space and time, evaluating distances requires inspection of n^2 pairs of points, and here steps should be made to be able to do this efficiently. The current data structures as such are usually able to handle the big data as well, but most likely specific procedures have to be developed that are able to address issues that are relatively novel (such as combining data in the space–time domain) or that have to address specific questions and problems, i.e., to select data from a big data set for a particular model application. A particular way ahead may be that classification of the data into multiple classes is done in the form of metadata. In such a way it is possible to make the big data of relevance in a wide range of practical applications. This would require an improved database structure, and in particular a very much adaptive spatial statistical analysis procedure.

Some authors have already pointed out the necessity of parallel and distributed programming for handling the big data sets in the general context or even in the geospatial context (Lee et al., 2014; Shekhar et al., 2012, 2014; Wang et al., 2013). Others have pointed out the usefulness of functional programming concepts or languages such as Haskell Domain-Specific Language (Mintchev, 2014), Map-reduce (Maitrey and Jha, 2015; Mohammed et al., 2014), Data Flow Graphs (Tran et al., 2012), or self-adjusting computation (Acar and Chen, 2013). However, there is a gap between the research works that advocate functional programming techniques but do not handle specifically geospatial data, and research works that focus on geospatial big data, but do not guarantee the absence of data races which are the races of different threads to gain access to the same data item in some shared memory Milewski (2009). The following sections discuss issues related to functional programming paradigms for big data streams and geospatial big data analytics in the context of big data modeling and structuring, and examines how geospatial data models and structures are adapted to big data.

4.1. Functional programming for big data streams

The main stumbling block for handling geospatial big data streams using parallel programming and the best reason for using functional programming is the concept of data races. A data race is a race between different threads that try to access the same data items, and relates to the notion of concurrency. Functional programming solves the problem of data races by strictly controlling the simultaneous access to mutable data. It has been predicted that data races will produce the "downfall of imperative programming" (Milewski, 2009). The main advantage of Haskell (or other functional programming languages like Closure, Lisp, ML, Scheme) for big data is its support of parallel computing and concurrency and the high performance of the most famous Haskell compiler (GHC: the Glasgow Haskell Compiler): "applications built with GHC enjoy solid multicore performance and can handle hundreds of thousands of concurrent network connections" (Mintchev, 2014).

Domain Specific Languages (DSL) provide a solution to the key challenge of big data by making the "multi-disciplinary collaboration as effective and productive as possible" and by offering the "required degree of flexibility and control" and a domain-specific development completed on time possibly by non-software developers (Mintchev, 2014). The functional programming languages map and reduce functions are the basis of the MapReduce programming model for processing big data sets by a distributed parallel algorithm. Maitrey and Jha (2015) states "MapReduce has emerged as the most popular computing paradigm for parallel, batch-style and analysis of large amount of data", especially since Google adopted it.

4.2. Geospatial big data analytics

We can classify the techniques used for spatial data mining from different points of view: the assumptions that these techniques pre-suppose and the "curse of dimensionality" that they exhibit or not. While parametric statistics assume some probability distribution function or some spatial distribution, non-parametric statistics only assume local smoothness. Functional analysis (e.g., wavelets) and homotopy continuation techniques assume only the continuity of the functions involved. The "curse of dimensionality" (Juditsky et al., 1995) states that the number of points needed to train machine-learning algorithms gets exponential with the dimension of the search space. Statistical and machine learning techniques and even statistics based dimensionality reduction algorithms do not lower the dimensionality of the problem in a deterministically exact way (Li et al., 2011). Thus, they exhibit the "curse of dimensionality" (Juditsky et al., 1995). On the contrary, homotopy continuation techniques are not subject to an exponential growth of the number of samplings (because they do not rely on training with points) and the uncertainty of the modeling does not explode from one dimension to the next one, as this has been shown in Musiige et al. (2013) and Musiige et al. (2011). This observation can be extended to other functional analysis techniques as wavelets (Juditsky et al., 1995).

As it can be conceived, in any attempt to process big spatial data streams in real time, one might be tempted to ask if tolerating soft errors could be feasible, and to what extent. This was addressed in Carbin et al. (2013). If we do not accept soft errors, then we need to rely on new High Performance Computing architectures to harness the parallelism necessary to process geospatial big data streams in real time (Carbin et al., 2013). However, in order to analyze and compare geospatial big data algorithms, we need benchmarks, so that the different algorithms can have a common evaluation basis. Big data benchmarks (Shekhar et al., 2014, 2012) have become one fundamental concept in studying geospatial big data.

Spatial databases research has intended to make query processing faster by designing spatial indexing methods, which partition the search space in tiles (possibly irregular), so that the average query time will be concentrated in one tile. The main challenge is to cleverly organize geospatial data in tiles through spatial data clustering that will define optimal tiles and efficient paths (space-filling curves) through these tiles, so that the access to n-dimensional data is done efficiently by referring to the location of the tile along that "space-filling curve". Several space-filling

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

curves have been proposed in the literature. However, the Hilbert space-filling curve has the advantage that the arc length between two consecutive tiles along the space-filling curve is constant (Ujang et al., 2014). The main use and value of geospatial big data is to dig useful information from geospatial big data sets (Wang et al., 2013; Wang and Yuan, 2013, 2014).

5. Data visualization and visual analytics

When several additional Vs are proposed in defining the big data (see Section 1 above), it is no surprise that the terms *visualization* and *visual analytics* are frequently mentioned. There are many strong reasons for this with the primary one being that some of our computational and statistical approaches do not scale – there is simply *too much* data (Keim et al., 2013; Shneiderman, 2014). Even smaller amounts of data in forms and tables are not really human readable, thus the interactive and exploratory visualization environments help at the very early stages in dealing with big data in making sense of what the data actually contains (Cook et al., 2012; Frankel and Reid, 2008; Hoffer, 2014). In other words, visualizations essentially enable humans to deal with big data where machines along might fall short. Conversely, visual analytics approaches acknowledge the human shortcomings as well and combine the powers of computational tools with powers of human visual sense-making (Choo and Park, 2013). Visualizations, therefore, are widely acknowledged as a part of analysis process (i.e., not only communication), in which we can explore the data, and build hypotheses during this process (Zhang et al., 2012). It is important to note that visualizations have been more commonly conceptualized as *communication* tools. While this is true and visualizations are important in communicating hypotheses, results and ideas, in the case of big data, we believe their role in *exploration* plays an even more important part. However, visualizations, while often offered as remedies to the shortcomings of the computational methods, may not scale either. Big data means many things to display; and it often results in very 'busy' displays, especially given the trend for multiple-linked view displays that are popular in visual analytics and big data applications. Such requirements can lead to information overload. It is important to note that human cognitive resources (such as the visual working memory that is critical in processing visual information; or spatial abilities which are critical for how well we can make sense of visualizations) are limited (e.g., Hegarty, 2011; Hegarty et al., 2012). Novel visualization designs are necessary, especially those informed by knowledge on human information processing, perception and cognition.

In terms of novel design and visualization paradigms, the field witnessed many alternative approaches that are constantly being proposed, even though many of them are not yet validated through empirical testing. Among these approaches, the most dominant one appears to be the *multiple-linked views*, in which approaches such as brushing and linking are used to allow the viewer to work with various visualizations at the same time (Bernasocchi et al., 2012). Alternatively, summarization, clustering and highlighting approaches have been proposed. Vision-inspired approaches, such as *focus + context* visualizations and *foveation* may provide interesting new opportunities as they attempt to reduce the information load and are becoming more relevant with the recent technological developments offering cheaper and better eye tracking solutions (Bektas and Çöltekin, 2012; Cockburn et al., 2008; Çöltekin, 2009). Other approaches have also been proposed in the literature taking advantage of technological developments such as *cloud computing*, *parallel processing*, *indexing and querying* for real time utilization (e.g., Chen et al., 2014; Hoffman, 2012; Liu et al., 2013; Lu et al., 2013).

Despite an influx of "new" solutions, we have also witnessed a rediscovery of a "not so new" system – through a strong coupling between geographic information systems (GIS) and big data. GIS is an unmatched and mature toolbox for *data science* as its abilities to process spatial and non-spatial (attribute) data (even when they are not perfectly structured) through computational as well as visual means. Some even called GIS and big data "two parts of a whole" (Deogawanka, 2014). Geographic information science and related domains such as remote sensing and geoinformatics have been dealing with large datasets for a relatively long time, before the term big data took momentum in science, popular culture and business (Çöltekin and Reichenbacher, 2011). With the advent of big data, other branches of geography have also shown great interest in utilizing big data for addressing social (human geography) and environmental (physical geography) questions (e.g., Crampton et al., 2013; Goodchild, 2013; Kitchin, 2013; Steed et al., 2013; Wood et al., 2013).

One of the challenges is to be able to make such geospatial big data accessible to end users so that it can be used to make real world decisions. Data visualization tools and techniques are therefore critical in providing windows into such rich data so that it can be analyzed and interrogated by researchers, policy and decision-makers and citizens alike. World Wide Web platforms, such as *geoportals*, provide an excellent means to deliver such services. Portals provide access to geospatial data, and there has been significant momentum in creating federated geoportal, which can access a window to a vast array of geospatial data sets. For example, the INSPIRE Geoportal (http://inspire-geoportal.ec.europa.eu/) provides access to 10,000s of geospatial metadata data records from across Europe. The INSPIRE Geoportal uses a visualization interface comprising of both a map window and folksonomy tag cloud to assist users in navigating this rich geospatial data resource.

There has been an explosion of data available and as our planet continues to experience significant rapid urbanization, there is an increasing need to access and visualize data which represents the dimensions of space and place (see Straumann et al. (2014) for a distinction of the terms space and place). On this note, the rise of *smart cities* has led to the instrumentation of cities with more real-time data and historical data being captured and visualized (Cheshire and Batty, 2012). Large-scale projects such as the Urban Big Data Centre (UBDC) (http://ubdc.ac.uk/) and the Australian Urban Research Infrastructure Network (AURIN) (http://aurin.org.au/) are endeavoring to develop "big data" visualization tools and techniques to support the realization of smart cities. The challenge is to be able to provide not only such data visualization interfaces to researchers, but also tools to support policy and decision makers, city planners and communities to be able to visually explore and analyze this data to make better decisions in collectively planning our cities.

For example, addressing a rapidly urbanizing Australia, AURIN has developed a geoportal where over 1800 datasets can be accessed and visualized. AURIN has deployed a federated data architecture which is metadata driven (Sinnott et al., 2015). There are over 6 billion data elements that urban researchers, government policy and decision makers can access via the AURIN portal (Pettit et al., 2014). This rich tapestry of "big data" across the domains of health, housing, transport, demographics, economics and other essential areas provides coverage across the major cities of Australia. Both *point in time* and *longitudinal data* are accessible via the AURIN portal. Other examples show how geospatial big data is being used in practice in connection with multi-user (collaborative) visualization environments in the context of geoportals for visualizing big data through work being undertaken in Europe through INSPIRE and in Australia through AURIN.

Another "cutting edge" research area is the visualization of networks, such as transportation networks (Cheshire and Batty, 2012). In the London Oyster card case mentioned before in Section 2, there is a need to be able to visualize individual and aggregated

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

Fig. 1. Oyster card data for London Tube and train stations, animated for a day using 10 min intervals (Created by Oliver O'Brien (http://oobrien.com/2013/03/londons-tidal-oyster-card-flow/) (via Batty, 2012).

travel journeys to provide further insights in the travel behaviors of commuters as they move through the city, and this, in turn, provides important information to transport planners in optimizing timetabling and responding to events. Fig. 1 illustrates a snapshot of a time sequence animation of a typical weekday travel based on the Oyster card data.

In addition, the real-time visualization of crowd-sourced big data from platforms such as Twitter can assist with, for example, disaster management responses. The visualization of real-time data streaming from these sources can provide emergency response team critical information on how to respond to events such as flood, fires and other natural and human induced disasters. Visualization platforms such as Ushahidi (http://www.ushahidi.com/) and Cognicity (http://cognicity.info/cognicity/) and the Peta Jakarta project (http://petajakarta.org/banjir/en/) are such crowd-sourced platforms. In the latter, citizens can tweet the reporting of floods particularly in the Monsoon season in the city of Jakarta, Indonesia.

6. Data mining and knowledge discovery

General big data analysis and traditional spatial data analysis and geo-processing methods and theories can all contribute to the development of geospatial big data analysis and processing. Statistical analysis, geo-computing, simulation and data mining methods and techniques can be used alone or together with other types of big data for discovering knowledge from geospatial big data. This section examines traditional knowledge discovery methods and resurgence of fractal analysis in dealing with geospatial big data.

6.1. Data mining and knowledge discovery

Knowledge discovery (KD) is concerned with mining and extracting meaningful patterns and relationships from large datasets that are valid, novel, useful and understandable (Miller and Hanz, 2009). The field emerged in response to the need for methods applicable to data that violate the assumptions of traditional statistics. KD methods typically emphasize generalization ability and predictive performance, which is particularly pertinent with spatio-temporal data because spatio-temporal datasets can provide rich information about how a process evolves over time. KD from spatio-temporal data enables us to create models that are able to predict future states of a process, so called the holy grail of science (Cressie and Wikle, 2011). KD encompasses a range of spatio-temporal data mining (STDM) tools and methodologies for carrying out a set of tasks.

Perhaps the most conceptually simple KD technique is *Association rule mining* (ARM), which involves searching for associations in a dataset where an event X tends to lead to an event Y, where X is the antecedent and Y is the consequent (Agrawal et al., 1993). In the context of spatio-temporal data, ARM entails searching for the occurrence of an event Y in the spatio-temporal neighborhood of another event X (Mennis and Liu, 2005). Shekhar et al. (2011) describe some of the types of patterns that may be present in spatio-temporal data.

More traditional data analysis methods also come under the remit of KD. *Regression* is viewed as a data mining technique, but in truth its use in the spatial sciences and time series analysis predates the field of data mining. Regression models for spatio-temporal data emerged from the cross-pollination of ideas from time series analysis, econometrics and the spatial sciences. For example, the space–time autoregressive integrated moving average (STARIMA) model (Pfeifer and Deutsch, 1980; Cheng et al., 2014a, 2014b), spatial panel data models (Elhorst, 2003), Bayesian hierarchical models (Cressie and Wikle, 2011), external-drift kriging (Van de Kassteele and Stein, 2006), space–time geostatistics (Stein et al., 1998; Heuvelink and Griffith, 2010) and image mining (Rajasekar et al., 2006), among others. An important form of regression is classification, in which the outputs are class labels. As larger and more granular datasets have become available in recent years, the limitations of traditional statistical approaches in capturing nonlinearity and heterogeneity have been exposed. Therefore, some scholars have looked to the machine learning (ML) community for alternatives, first to artificial neural networks (ANNs) in the 1990s and more recently to kernel methods. Perhaps the most well-known and broadly successful machine learning (ML) method for classification and regression is the support vector machine (SVM), which uses kernels to carry out nonlinear regression or classification (Kanevski et al., 2009; Haworth et al., 2014). Recently, the Random Forest (RF) which is a method that combines multiple decision trees through bootstrapping has also gained popularity for classification (Cutler et al., 2007).

Other KD tasks include *anomaly detection*, also known as outlier detection, and *clustering*. Anomaly detection involves the identifi-

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

cation of events or patterns in data different from what one would expect. Anomaly detection is inherently challenging as it requires a clear definition of what is considered to be normal and abnormal. In spatio-temporal processes, these definitions may evolve and change over time (Chandola et al., 2009). Accounting for these changes in different spatio-temporal processes is a key research challenge. *Clustering* is a form of unsupervised learning, which involves uncovering hidden structure in a dataset about which we know little. Clustering has wide applications in the spatial sciences, for example in geodemographic classification (Vickers and Rees, 2007) and hotspot detection (Nakaya and Yano, 2010). Although spatial clustering methods are well developed, spatio-temporal clustering (STC) is still an emerging research frontier. STC methods that are gaining popularity include ST-DBSCAN (Birant and Kut, 2007) and space–time scan statistics (STSS) (Kulldorff et al., 2005; Cheng and Adepeju, 2014).

The impact of ML methods on KD and STDM has been significant. ML methods are generally effective in tackling nonlinearity in spatial data, and can be modified to deal with the multi-scale issue and heterogeneity (Foresti et al., 2011). However, in many cases (especially kernel methods) their initial computation is expensive if the number of data samples is large. Furthermore, if the statistical properties of a space–time series evolve over time, models have to be retrained to reflect this. In the big data age where the ability to apply methods to real time data streams is paramount, new ways of training traditional algorithms are needed. In ML, online learning is used (Castro-Neto et al., 2009) and these types of approaches need to be integrated with more traditional spatio-temporal analysis techniques. Parallel and grid computation can also be used to improve the performance of KD methods (Harris et al., 2010). However, some issues cannot be solved with only improvements in computational efficiency. For example, the principal problem in STC is to model how clusters emerge, change, move and dissipate/disappear in time. This can be achieved retrospectively but is very difficult to quantify in time critical applications. At what point does a cluster of crimes become a hotspot? Current methodologies within KD and STDM are generally designed to analyze historical datasets but cannot adequately deal with evolving properties of space–time data.

6.2. Fractals emerged from big data

Big data show incredible fractal structure, in which there are far more small things than large ones (Jiang, 2015a; Jiang and Miao, 2015). There are several reasons for the emerging fractals. First, big data are usually emerged from the bottom up or are contributed by diverse individuals, e.g., location based social media data, so they are very diverse and heterogeneous. Second, big data are defined at very high spatial and temporal scales, which enable us to observe fractal structure and nonlinear dynamics more easily. Third, big data due to the size are more likely to capture a true picture of reality, and reality is no doubt fractal (Bak, 1996; Mandelbrot and Hudson, 2004). In these aspects, the emerging fractals differ from fractals seen on small data such as fractal cities (Batty and Longley, 1994). We argue that fractal geometry (Mandelbrot, 1982), or complexity science in general, should be adopted for big data analytics and visualization, and for developing new insights into big data.

Let us examine a working example to see how fractals are generated from tweets location data of the entire world (Jiang, 2015a; Jiang and Miao, 2015). The data were sliced at different intervals, and in an accumulative manner, which means that locations at t_1 are included in locations at t_2. For each sliced location dataset, we built a triangulated irregular network (TIN), and merged those small TIN edges (smaller than the mean of all the TIN edges) as individual patches, which are referred as natural cities. Eventually, thousands of the natural cities are emerged from the tweets locations. Fig. 2 shows two natural cities near Chicago and New York at the four time instants, and they are put in comparison with the generative fractal – Koch snowflake. The two natural cities look

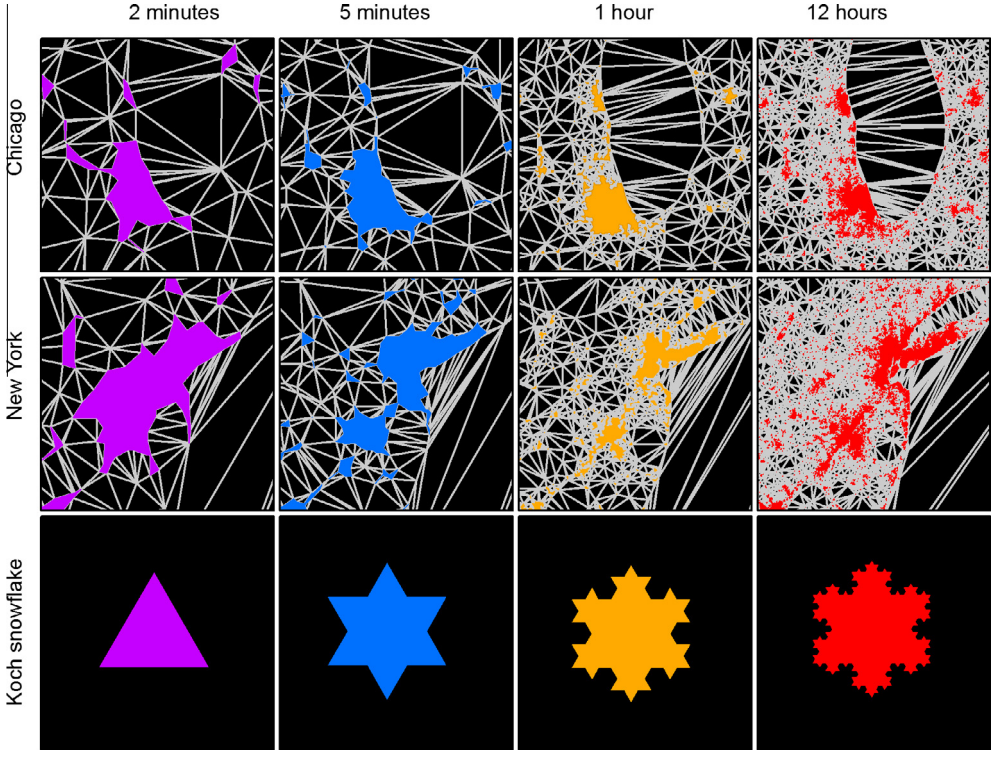

Fig. 2. Fractals emerged from big data look very much like the generative fractal – Koch snowflake (Jiang, 2015a).

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

very much like the snowflake, both sharing (1) the scaling pattern of far more small things than large ones, and (2) irregular shapes or boundaries. On the other hand, the natural cities look a bit different from the snowflake; the natural cities are developed from some irregular patches, which become further fragmented, whereas the snowflake and its growth come from the regular triangle with a strict scaling ratio 1/3. In other words, the former is called statistical fractal, being statistically self-similar, while the latter called strict fractal, being strictly self-similar.

What the example illustrated are not only fractals or natural cities generated from big data, but also a new, relaxed definition of fractals. A set or pattern is fractal if there are far more small things than large ones in it, or the scaling pattern of far more small things than large ones recurs multiple times (Jiang, 2015b; Jiang and Yin, 2014). This new definition is in fact developed from head/-tail breaks (Jiang, 2013) as a classification scheme for data with a heavy tailed distribution. The complexity of fractals can be measured by the head/tail breaks induced index – ht-index: the higher the ht-index, the more complex the fractal. Conventionally, complexity was captured by fractal dimension (Mandelbrot, 1982), thus ht-index being an alternative index to fractal dimension. In comparison to conventional definitions of fractal, the new definition is much more intuitive and easier to understand, so that anyone can rely on it to see fractals. For example, society is fractal because there are far more poor people than rich people, or far more ordinary people than extraordinary people (Zipf, 1949). It should be noted that the head/tail breaks method is not only for data classification, with which both the number of classes and class intervals are automatically determined, but also an efficient and effective visualization tool (Jiang, 2015b). Fig. 3 presents an example of visualization using the head/tail breaks; only head part of the whole is shown to the right panel, but the part is self-similar to the whole, and thus the part reflects the whole.

Fractal geometry represents a new way of thinking for geospatial analysis, and this is particularly truly for big data analytics. Despite that Euclidean geometry has thousands of years of history, and serves as the foundation of geospatial technologies, the essence of geography (both physical and human) is fractal. We therefore must adopt fractal methods for developing new insights into big data. This fractal thinking is in line with Paretian thinking (Jiang, 2015b), which is in contrast to conventional Gaussian thinking. Statistically speaking, there are far more small things than large ones, rather than more or less similar things. Given the scaling pattern or the heavy tailed distribution, the head/tail breaks provide an effective means to derive an inherent hierarchy of complex systems. This is the first step towards an understanding of geographic phenomena, i.e., to recognize fractal structure of geographic systems. The next step is to further develop, through simulations, an understanding of processes as to why the fractal structures exist. In this regard, a set of complexity modeling tools have been developed such as cellular automata, agent-based modeling, and the sand pile model.

7. Challenges, open issues and future directions

This section presents the challenges and open issues based on the reviews in Sections 2–6 and outlines some research directions in the three focus areas of ISPRS TC II (http://www2.isprs.org/commissions/comm2.html).

7.1. Efficient representation and modeling for geospatial big data

The main problem that spatial algorithms face in the context of real-time big data handling is that they cannot wait until all the data are known, as it is the case in two of the major classes of spatial algorithms: the divide-and-conquer or line/plane-sweep algorithms. Even though incremental algorithms are well-suited for handling changing data sets, they are not well-suited for handling streams that cannot fit in the main memory of the computer. Therefore, a new class of spatial algorithms has started to be designed and developed: spatial data streaming algorithms (see e.g., Hoffmann et al., 2007; Sharifzadeh and Shahabi, 2009). However, any real-time streaming algorithms must read an input stream, process it, and write the output stream in real-time. Among all the tasks that have to be performed on computers or digital circuits, the tasks that are the slowest, i.e., communication over a network, must be minimized in order to satisfy the real-time constraint of big data processing. Another implication of the constraint of real-time processing of big spatial data is to minimize the amount of disk

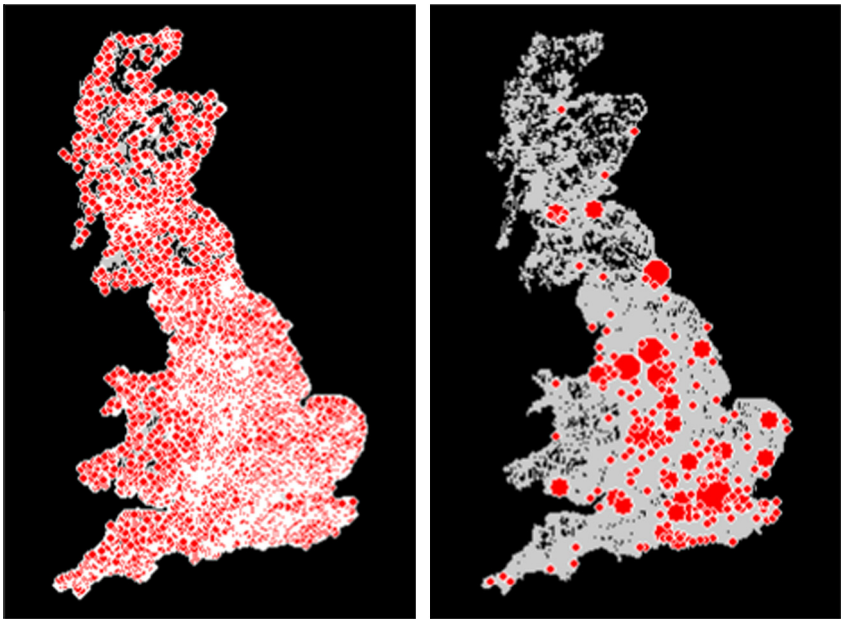

Fig. 3. Head/tail breaks as an effective visualization tool. (Note: 16,000 natural cities (to the left) in UK generated from a half million of points of interest look like a hairball, but their top 4 hierarchical levels clearly show a scaling structure to the right.)

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

input/output, which is the second slowest task after communications over a network. For this purpose, one needs to focus on the most compact data structures that will store spatial data using the smallest amount of memory. Finally, as spatial data tend to be more complex than non-spatial data, the total CPU running time is not negligible with respect to the disk input/output time. Thus, the parallelization of the spatial algorithms will bring a non-negligible speed-up to the spatial algorithm. This means the spatial algorithms must be distributed, parallelized and use the most compact spatial data structures so that the exchange of information over a network and between the disk and the main memory are minimized and the CPU running time is also minimized. This imposes a distributed parallel architecture where streams of data are processed in real time on each unit controlling a sensor in order to transmit over the network only summary statistics and required results for the other nodes in the computing environment. Such summary statistics that are relevant for sensors are intervals, in particular the new measurements that change the intervals of measurements or the required results. Interval analysis is a well-known method for computing bounds of a function, being given bounds on the variables of that function, (see Moore et al. (2009, p. 223) for an introduction to interval analysis). The basic mathematical object in interval analysis is the interval instead of the variable. The operators need to be redefined to operate on intervals instead of real variables. This leads to interval arithmetic. In the same way, the most usual mathematical functions are redefined by an interval equivalent. Interval analysis allows one to certify computations on intervals by providing bounds on the results. The uncertainty of each measure can be represented using an interval defined by either a lower bound and a higher bound or a midpoint value and a radius.

Furthermore, interval analysis is in conjunction with functional analysis, the most-well-suited framework to model the uncertainty of geospatial big data. First, as observed by Juditsky et al. (1995), functional analysis methods like wavelets do not suffer from the curse of dimensionality that affects machine-learning as well as parametric statistics (including multivariate statistics). Second, interval analysis allows one to model the uncertainty of the input variables (like sensor observations) and the corresponding uncertainty of the functions that are evaluated on these variables. The key challenge and open issue is to bridge the gap between the machine learning and functional analysis communities by convincing the communities working on and with geospatial big data to use functional analysis and interval analysis with a functional programming language. Unfortunately, despite functional programming has made its way into other programming language paradigms, functional analysis combined with interval analysis has not been so successful in making its way in applications of big spatial data.

Four initial observations can be seen as the base for exploring further research directions. Firstly, functional analysis methods like wavelets, homotopy continuation and interval analysis are much better suited than parametric statistical methods to cope with the curse of dimensionality inherent in big spatial data that includes many dimensions (each functionally independent physical value measured corresponds to one dimension and each one of the coordinate systems component x, y and z corresponds also to one dimension). Secondly, pure functional programming is very well suited for handling functions because pure functional programming does not have (unlike impure functional programming and other programming paradigms) side effects in pure functions, and functions are one of the two most fundamental concepts in any functional programming language: functions and data types (see the seminal paper of Hughes (1989)). Thirdly, the main challenge in spatial data handling, which is to certify the topological relationships and the uncertainty of any spatial data modeling or decision-making by determining the uncertainties of all input variables, can be solved using interval analysis. Finally, the lazy functional programming paradigm, which postpones the evaluation of expressions until these expressions need to be computed in order to compute the final result (Hughes, 1989), is very well suited for fractals due to its ability to represent fractal recurrences and to compute only the fractal components needed to compute the final result.

One future research focus in his area is to produce a locally distributed stream sensing, processing and telecommunicating paradigm implemented using:

- new functional specification methods derived from ontologies, ontology mappings and their gluing into ontology categories and their charts;
- functional analysis methods (decompositions of streams with interval enclosures of wavelets, mathematical modeling in higher dimensional measure spaces from interval valued homotopies in lower dimensional measure spaces);
- pure functional programming languages (see haskell.org committee, 2015; Hughes, 1989) with:
  - visualization libraries (like Haskell and its OpenGL bindings and many specialized visualization libraries),
  - parallelization and concurrency libraries (like the distributed MapReduce framework Holumbus or OpenCL and OpenMP wrappers),
  - cloud computing libraries;
- functional programming inspired parallelisation and concurrency techniques.

The architecture of any system based on this paradigm can be considered a fractal. Every sensor controlling unit is responsible for collecting the big data streams, computing the statistics or any other desired result, generating the triggers that will automatically update any computed result, visualization or decision making and transmitting it to a lower-resolution data collecting node. Each local data-collecting node is responsible for storing the streams of results and providing the desired visualization and assembling the partial decision taking elements into a summarized decision taking from the neighboring sensor controlling units. Each regional data collecting node is responsible for storing the streams of results and providing the desired visualization and assembling the partial decision taking elements into a summarized decision taking from the neighboring local data collecting nodes.

The other future research focus on new processing algorithms to handle large volumes of data through use of functional programming languages is to design new streaming algorithms that:

- use the provably most compact geometric topology data structure that encodes all the ramifications (i.e. the singular points of the skeletons of the objects),
- use CPU and GPU parallelization to harness the computing facilities wherever these lie (locally),
- use interval analysis to:
  - represent uncertainty of streams of spatial data (Dilo et al., 2007), and
  - to automatically generate the triggers that will react automatically once an input value in a stream will make a change in any computed result, visualization or decision making;
- use wavelet decompositions for handling signals acquired by sensors,
- use fractals for handling spatial data that has components whose dimension is not an integer (the Hausdorff-Besicowitz dimension of a fractal is not an integer, and thus, fractals can be used to model phenomena that exhibit self-similarity),
- use interval valued homotopies to model or reconstruct functions in higher dimensional measure spaces from measure-

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

ments and reconstructions of functions from lower dimensional measure spaces, reconstructing therefore higher dimensional measure spaces one dimension at a time,
- use categories to represent the functional dependencies between data variables and any computed result or visualization or decision taking (like data types and functions in the Hask category), and a category based Domain Specific Language computing library (like docon, see Mechveliani (2001)).

7.2. Analyzing, mining and visualizing geospatial big data for decision-support

Spatial statistical methods are in principle able to also include non-spatial big data. A typical example is the use of co-variables in a spatial interpolation procedure (Van de Kassteele and Stein, 2006), or in a spatial point process modeling analysis. Such analyses commonly rely on assumptions, such as normality, independence, absence of noise in the explanatory data. With the advancements of big data, however, on the one hand the quality of the big data can be doubted, whereas on the other hand the speed of calculations is seriously affected. Hence, the procedures are at the moment not ready for the purpose, and serious preprocessing has to be done, for which the tools are not readily available. Some of the methods have been developed in the past, such as possibilities to use explanatory variables of a changing quality, but these are often rather complicated to use, and require substantial work to make them available for the purpose. As an example, one may think of Bayesian procedures, where prior distributions can be included into a likelihood function. These methods have shown an enormous power in spatial studies but their application on an automatic basis for big data may not be so easy and transparent to develop, implement and apply.

The main challenges in KD and STDM can be summarized around the first four *Vs* described in Section 1.

Volumes of spatio-temporal data are ever increasing and some argue that traditional transactional database structures are becoming outmoded. Although this point is open to debate, research is taking place into how to best use new data storage and query architectures to deal with spatio-temporal data. For example, recent studies have used MapReduce programming models to parallelise data processing (Tan et al., 2012). Despite this, traditional SQL based spatial databases, such as PostGIS and Oracle Spatial, remain dominant in academia. If this trend continues, there is a risk that academic research will become disconnected from industry.

Despite significant progress, challenges remain in developing predictive algorithms that can deal with the *velocity* of data arriving in real time. Geography has been a big data discipline since long before the term arose, dealing with large and complex problems like weather and climate modeling. Hence, geography has traditionally been forward thinking regarding the development of algorithms for dealing with large and complex datasets in a timely fashion. The discourse on parallel computing in geography began in the 1990s, leading to the emergence of the subfield of geocomputation (Cheng et al., 2012). However, parallelism is still far from the common practice despite enabling hardware being available in desktop computers. Most recent work has focussed on parallelising all or part of existing algorithms to improve computational performance. For example, Guan and Clarke (2010) developed a parallel raster processing library for use in cellular automata, and Guan et al. (2011) developed algorithms to parallelize elements of Kriging interpolation. Libraries for parallel geocomputation are also now beginning to emerge in open source software environments such as R (Harris et al., 2010). Building on this, a greater focus needs to be placed on developing algorithms that are parallel in nature and can harness all types of parallelism. This is what Turton and Openshaw (1998, p. 1842) termed "Thinking in Parallel" in 1998, but is yet to be fully adopted in the research community.

Variety presents itself as an opportunity to the research community. We now have potentially many data sources of different types that can be used to analyze (and re-analyze) a diverse range of spatio-temporal processes. Traditional data providers such as governments, national mapping agencies and transport authorities are now complemented by new data sources described in Section 2. However, an important and often overlooked problems of KD in the spatial sciences is data *veracity*. Big data provides unprecedented volumes of data about a broad range of human activities and physical processes. However, they are often collected on a fairly ad-hoc basis when compared with traditional data sources, and usually must be repurposed to fulfil research objectives. Initiatives such as OpenStreetMap have proven that crowd sourced data can compete with official sources in this regard, but issues of sample representativeness and data collection design (or lack thereof) are still a concern. A good example is Twitter, which has gained popularity in research communities recently, but under represents certain groups, including the elderly and some ethnic minorities, and has an uneven spatio-temporal distribution. Such datasets clearly have potential value alongside traditional demographic data such as censuses and cross-sectional surveys (Longley et al., 2015). However, there is a trade-off between the spatial and temporal granularity offered by these data and the certainty associated with any conclusions drawn from them. Personal mobility data is another example; smartphone apps collect and store vast quantities of such data which have considerable research potential, but this cannot be fully realized without proper validation and a clear understanding of potential biases in the user group. This is coupled with difficulties in inferring activities from raw, unlabelled mobility data (Bolbol et al., 2011).

Sampling is not a legitimate concept in the big data era, as argued by Jiang and Miao (2015). Big data tends to take all or a large amount, and this data characteristic make big data different from small data which are often sampled. Surely, social media data are oriented towards younger generations or those who have access to Internet and social media, and not everyone has Internet access, in particular in developing countries. However, the large data volumes enable big data to capture the true picture of all social media users, despite the fact that not all people are involved in social media. All social media users can be a good proxy for studying real human activities on the earth surface. In this regard, social media data are like maps, which do not represent everything on the earth surface, but can be a good proxy of earth surface. Big data calls for new ways of thinking beyond conventional thinking (Jiang, 2015b; Mayer-Schönberger and Cukier, 2013). For example, we tend to examine data quality as we did in the small data era. Massive data should tolerate messiness. For navigation purposes, we must make OSM data as precise and accurate as possible, in other words, the more precise and the more accurate, the better. However, to examine whether there are far more small street blocks than large ones, we do not need very good quality data, but need massive data with messiness instead.

Nonetheless, visualizing big data has some additional challenges; for example, visual analytics solutions themselves may not scale: we need to consider how to deal with *information overload* on the viewer if we show too many things at the same time (Choo and Park, 2013; Ruff, 2002). Applications in a large-scale geoportal like AURIN have revealed there are significant challenges in being able to visualize large multi-dimensional geospatial datasets via a browser. For this, thousands of years of cartographic expertise offers sound advice. We should carefully generalize, e.g., emphasize the important while removing the unimportant, group the information both thematically and perceptually, and pay attention to visual

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

hierarchy when we design displays. We should remember linking the good cartographic design principles to modern interaction design paradigms (MacEachren et al., 2008; Schnürer et al., 2014; Slocum et al., 2008). Furthermore, researchers in the cartography and geovisualization domain have taken a strong interest in cognitive and usability issues and much progress has been made to understand how human capacity can enhance or limit our experiences with visual displays (Çöltekin et al., 2010; Knapp, 1995; Montello, 2002; Roth, 2013; Slocum et al., 2001).

While the *smart city* and *big data* are hot topics of today (Batty, 2012), there is also the challenge of how to visualize the error and uncertainty inherent with big data sets such as crowd-sourced datasets and smart card data (Cheshire and Batty, 2012). In the realm of geospatial big data, there is also a need to effectively visualize other types of massive data sets, such as LIDAR data or very large collections of remote-sensing data. Visualization platforms such as PointCloudViz (http://www.pointcloudviz.com/), or Online LIDAR point cloud viewer (http://lidarview.com/) specialize on Lidar data visualization. Similarly, there are efforts in organizing remotely sensed imagery (Ma et al., 2014; Marshall and Boshuizen, 2013).

Spatial statistics has always had good opportunities for visualizing spatial and spatio-temporal data, including the uncertainties (see Rulinda et al. (2013) for an excellent example in the space–time domain). Their opportunities are still present also for geospatial big data. In the past, such methods have also been applied to non-spatial data, and there is apparently not a real problem to extend them towards big data. A critical issue in all of spatial (or spatio-temporal) data is their relying on coordinates. Hence, also for non-spatial data there must be an opportunity to assign coordinates, or equivalent, to the non-spatial data. Successes in the past have typically considered class memberships to serve that purpose, bringing the non-spatial data towards the feature space (Dilo et al., 2007).

7.3. Quality assessment of geospatial big data from new sources

Dealing with veracity in a scalable and timely manner has been identified as a substantial challenge (Saha and Srivastava, 2014). Behind this challenge is the lack of thorough knowledge of the data semantics when data that is commonly called "dirty" (e.g., Chiang and Miller, 2008) is collected and combined in an ad-hoc manner. Accordingly, big data analytics has given up the closed-world assumption in favor of learning incrementally under an open-world assumption. Thus, big data quality assessment has been characterized by the three stages of discovering rules of data semantics, checking for inconsistencies based on currently known rules, and repairing data near real-time (Saha and Srivastava, 2014).

Geospatial big data collection, especially sensors' data, is being captured in an automatic and unprecedented way, which poses new opportunities and challenges. This in principle causes problems for a statistical analysis, where statistical considerations such as optimal design or model based sampling are critical to make valid statements. Big spatial data may be either largely irrelevant, as they are collected automatically and hence much uninteresting repetition may occur, whereas big spatial data may be of a highly varying quality. They may be precise, but too precise to be of use, they may be too abundant as the same object is over sampled, they may be of a very poor quality as data may be at the nominal scale, where ratio-scale data might have been collected elsewhere, or they may be inadequate for the specific purpose. At the moment, no adequate procedures seem to be implemented to be able to harmonize the quality of spatial data for specific purposes. This will require a significant effort in order to develop feasible solutions.

On the positive side, the large number of data sources allows us to acquire a "complete picture" of a spatial situation, also including its dynamics. This is even more so, when different sensor data are integrated and fused, allowing to also eliciting more than one aspect or feature of an object (Ebert et al., 2009). Due to the potential high redundancy, it is possible to identify errors or blunders in the data and achieve higher accuracies even though an individual sensor has a limited measuring quality.

As often no explicit semantics (or manual annotation of semantics) is given, automatic processes are required to reveal it. This refers both to the object and its features, but also to the temporal characteristics.

Crowd-sourced data sets are often related to very detailed, i.e., large-scale phenomena. However, the level of granularity (in space, time and semantics) is not always known. Hence, it is a challenge to determine this granularity level from the sensor data. One option is to include explicit metadata about the sensor in terms of a self-description. Another option is to infer the scale and granularity level by automatically relating it to other sources of known granularity. This involves matching techniques at both the geometric and semantic levels.

In traditional sensors the quality (e.g., geometric accuracy) of the acquired data is given. This is not necessarily the case with new data sources such as VGI data. As this data source is often user centred, questions of reputation and trust have to be included to evaluate the quality.

8. Conclusions

This study reviews a variety of geospatial theory and methods used for traditional data but that can be extended to handle geospatial big data. While there is no standard definition of big data, it can be 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. Given these unique characteristics, traditional data handling approaches and methods are inadequate and the following areas were identified as in need for further development and research in the discipline:

- The development of new spatial indexing and algorithms to handle real-time, streaming data and to support topology for real-time analytics.
- The development of conceptual and methodological approaches to move big data from descriptive and correlation research and applications to ones that explore casual and explanatory relationships.
- The development of efficient methods to display data integrated in the three dimensions of geographic and one dimension of continuous time. There is a strong need in understanding human capacity to deal with visual information and identifying which visualization type is a good fit for the task at hand, and the target user group. Furthermore, interdisciplinary studies and communication is important. The advances in scientific visualization and information visualization are both beneficial to geographic visualization; but geographic visualization has also a lot to offer to other domains. Novel visualization paradigms, especially developed for big data tend to be information-rich (thus complex); therefore, we find that highlighting and summarizing approaches should be further investigated. Additionally, and in relation to managing complex visualization displays, technology research in terms of level of detail management remains important.
- The development of novel approaches for error propagation so as to effectively assess data quality requires. The challenge is not only the handling the many different types of data for real-time analytics, but rather the ad-hoc combination of data

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

streams in real-time, which may include the capture of the "whole" picture (or "complete population") instead of sampling a small portion of the whole population. In this case quick assessments are preferable that may come out of varying the input data and simulating variability.

The paper also highlighted other general conceptual and practical issues. The relation between spatial statistics and semantics and ontologies has been identified in the past, but requires further elaboration. Ontologies have been identified, for example, for dunes and beaches in studies around 2005, whereas more slum ontologies have been developed. The role of spatial statistics was related to the scale, the environment and the characterization of specific variables. In particular, aspects of scale are important.

Privacy and security are equally important and key concerns especially in geospatial big data handling, which may lead us into a "naked future" if not properly addressed. They are an essential part of geospatial big data management, but are not covered here given the focus on this paper on data handling methods.

Big data presents both challenges and opportunities. This paper outlines some of these challenges from a technical and conceptual perspective, and also provides priority areas that need to be addressed in the future. Once big data research evolves and matures, the opportunities from leveraging big data for overall societal management and decision-making become enormous.

Authors contribution statement

Regardless of the listed order of the authors, all authors contributed equally by participating in discussions, writing sections, revising corresponding sections and providing revision comments on the entire paper.

Acknowledgments

The partial support of this study was funded by the National Science and Engineering Research Council of Canada (NSERC) Discover Grants awarded separately to the first and second authors. The authors thank Anthony Lee (Spatial Analysis and Modeling Laboratory, Simon Fraser University, Department of Geography) for assistance in compiling the reference database and formatting the citations.

References

Acar, U.A., Chen, Y., 2013. Streaming big data with self-adjusting computation. In: Proceedings of the 2013 workshop on data driven functional programming – DDFP '13, pp. 15–18, doi:http://dx.doi.org/10.1145/2429376.2429382.

Ackermann, F., 1994. Digital Elevation Models – Techniques and Applications, Quality Standards, Development. IAPRS, 30/4(Commission IV), pp. 421–432.

Agrawal, R., Imieliński, T., Swami, A., 199). Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 22(2), pp. 207–216, doi: http://dx.doi.org/10.1145/170035.170072.

Anthony, D., Henderson, T., Kotz, D., 2007. Privacy in location-aware computing environments. IEEE Pervasive Comput. 6 (4), 64–72.

Bak, P., 1996. How Nature Works: The Science of Self-Organised Critically, first ed. Copernicus.

Batty, M., 2012. Smart cities, big data. Environ. Plan. B: Plan. Des. 39 (2), 191–193. http://dx.doi.org/10.1068/b3902ed.

Batty, M., Longley, P.A., 1994. Fractal Cities: A Geometry of Form and Function. Academic Press, London.

Bektas, K., Çöltekin, A., 2012. Area of interest based interaction and geovisualization with WebGL. In: Proceedings of The Graphical Web Conference 2012.

Bernasocchi, M., Çöltekin, A., Gruber, S., 2012. An open source geovisual analytics toolbox for multivariate spatio-temporal data for environmental change modeling. ISPRS Ann. Photogram., Remote Sensing Spatial Inform. Sci. 1–2 (2), 123–128.

Birant, D., Kut, A., 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data Knowl. Eng., Intelligent Data Mining 60, 208–221. http://dx.doi.org/10.1016/j.datak.2006.01.013.

Bolbol, A., Cheng, T., Haworth, J., 2011. Using a moving window SVMs classification to infer travel mode from GPS data. In: Proceedings of the 11th International Conference on GeoComputation, pp. 262–270.

Carbin, M., Misailović, S., Rinard, M.C., 2013. Verifying Quantitative Reliability for Programs that Execute on Unreliable Hardware. ACM SIGPLAN NOTICES – OOPSLA '13, 48(10), pp. 33–52, doi:http://dx.doi.org/10.1145/2544173.2509546.

Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., Han, L.D., 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Syst. Appl. 36, 6164–6173. http://dx.doi.org/10.1016/j.eswa.2008.07.069.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: a survey. ACM Comput. Surveys 41 (3), 1–58. http://dx.doi.org/10.1145/1541880.1541882.

Chen, M., Mao, S., Liu, Y., 2014. Big data: a survey. Mobile Networks Appl. 19 (2), 171–209. http://dx.doi.org/10.1007/s11036-013-0489-0.

Cheng, T., Adepeju, M., 2014. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. PLOS ONE 9 (6), e100465.

Cheng, T., Haworth, J., Anbaroglu, B., Tanaksaranond, G., Wang, J., 2014a. Spatiotemporal data mining. In: Fischer, M.M., Nijkamp, P. (Eds.), Handbook of Regional Science. Springer, Berlin Heidelberg, pp. 1173–1193.

Cheng, T., Haworth, J., Manley, E., 2012. Advances in geocomputation (1996–2011). Comput., Environ. Urban Syst. 36 (6), 481–487. http://dx.doi.org/10.1016/j.compenvurbsys.2012.10.002.

Cheshire, J., Batty, M., 2012. Visualisation tools for understanding big data. Environ. Plan. B: Plan. Des. 39 (3), 413–415.

Cheng, T., Wang, J., Haworth, J., Heydecker, B., Chow, A., 2014b. A dynamic spatial weight matrix and localized space-time autoregressive integrated moving average for network modeling, geographical analysis. Geogr. Anal. 46, 75–97. http://dx.doi.org/10.1111/gean.12026.

Chiang, F., Miller, R.J., 2008. Discovering data quality rules. Proc. VLDB Endowment 1 (1), 1166–1177. http://dx.doi.org/10.14778/1453856.1453980.

Choo, J., Park, H., 2013. Customizing computational methods for visual analytics with big data. IEEE Comput. Graphics Appl. 33 (4), 22–28.

Cockburn, A., Karlson, A., Bederson, B.B., 2008. A review of overview+detail, zooming, and focus+context interfaces. ACM Comput. Surveys 41 (1), 1–31. http://dx.doi.org/10.1145/1456650.1456652.

Çöltekin, A., 2009. Space-variant image coding for stereoscopic media. In: Picture Coding Symposium, 2009 (PCS 2009), pp. 1–4, doi:http://dx.doi.org/10.1109/PCS.2009.5167396.

Çöltekin, A., Fabrikant, S.I., Lacayo, M., 2010. Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. Int. J. Geogr. Inform. Sci. 24 (10), 1559–1575. http://dx.doi.org/10.1080/13658816.2010.511718.

Çöltekin, A., Reichenbacher, T., 2011. High quality geographic services and bandwidth limitations. Future Internet 3 (4), 379–396. http://dx.doi.org/10.3390/fi3040379.

Cook, K., Grinstein, G., Whiting, M., Cooper, M., Havig, P., Liggett, K., et al., 2012. VAST challenge 2012: visual analytics for big data. In: Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 251–255, doi:http://dx.doi.org/10.1109/VAST.2012.6400529.

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. Inform. Sci. 40 (2), 130–139. http://dx.doi.org/10.1080/15230406.2013.777137.

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. John Wiley & Sons.

Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

Dasgupta, A., 2013. Big data: the future is in analytics. Geospatial World.

Deogawanka, S., 2014. Empowering GIS with Big Data. <http://www.gislounge.com/empowering-gis-big-data/> (retrieved May 4, 2015).

Dilo, A., By, R.A.d., Stein, A., 2007. A system of types and operators for handling vague spatial objects. Int. J. Geogr. Inform. Sci. 21 (4), 397–426. http://dx.doi.org/10.1080/13658810601037096.

Dittrich, A., Lucas, C., 2013. A step towards real-time detection and localization of disaster events based on tweets. In: Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management, pp. 868–872.

Duckham, M., 2013. Decentralized Spatial Computing. Springer, Heidelberg.

Duckham, M., Kulik, L., 2006. Location privacy and location-aware computing. Dyn. Mobile GIS: Invest. Change Space Time 3, 34–51.

Ebert, A., Kerle, N., Stein, A., 2009. Urban social vulnerability assessment with physical proxies and spatial metrics derived from air- and spaceborne imagery and GIS data. Nat. Hazards 48 (2), 275–294. http://dx.doi.org/10.1007/s11069-008-9264-0.

Elhorst, J.P., 2003. Specification and estimation of spatial panel data models. Int. Regional Sci. Rev. 26 (3), 244–268. http://dx.doi.org/10.1177/0160017603253791.

Evans, M.R., Oliver, D., Zhou, X., Shekhar, S., 2014. Spatial big data: case studies on volume, velocity, and variety. In: Karimi, H.A. (Ed.), Big Data: Techniques and Technologies in Geoinformatics. CRC Press, pp. 149–176.

Fitzner, D., Sester, M., Haberlandt, U., Rabiei, E., 2013. Rainfall estimation with a geosensor network of cars – theoretical considerations and first results. Photogr.-Fernerkundung-Geoinformation 2013 (2), 93–103. http://dx.doi.org/10.1127/1432-8364/2013/0161.

Foresti, L., Tuia, D., Kanevski, M., Pozdnoukhov, A., 2011. Learning wind fields with multiple kernels. Stochastic Environ. Res. Risk Assess. 25 (1), 51–66. http://dx.doi.org/10.1007/s00477-010-0405-0.

Frank, A.U., 2001. Tiers of ontology and consistency constraints in geographical information systems. Int. J. Geogr. Inform. Sci. 15 (7), 667–678.
Frank, A.U., 2007. Data quality ontology: an ontology for imperfect knowledge. In: Winter, S., Duckham, M., Kulik, L., Kuipers, B. (Eds.), Spatial Information Theory. Springer, Berlin Heidelberg, pp. 406–420.
Frankel, F., Reid, R., 2008. Big data: distilling meaning from data. Nature 455 (7209). http://dx.doi.org/10.1038/455030a, 30–30.
Frommberger, L., Schmid, F., Cai, C., 2013. Micro-mapping with smartphones for monitoring agricultural development. In: Proceedings of the 3rd ACM Symposium on Computing for Development, p. 46.
Fuchs, G., Andrienko, N., Andrienko, G., Bothe, S., Stange, H., 2013. Tracing the German centennial flood in the stream of tweets: first lessons learned. In: Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, pp. 31–38.
Gomes, L., 2014. Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts, October 20, 2014 <http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts> (retrieved May 4, 2015).
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221. http://dx.doi.org/10.1007/s10708-007-9111-y.
Goodchild, M.F., 2013. The quality of big (geo)data. Dialogues Human Geogr. 3 (3), 280–284. http://dx.doi.org/10.1177/2043820613513392.
Goodchild, M.F., Glennon, A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. Int. J. Digital Earth 3 (3), 231–241.
Guan, Q., Clarke, K.C., 2010. A general-purpose parallel raster processing programming library test application using a geographic cellular automata model. Int. J. Geogr. Inform. Sci. 24 (5), 695–722. http://dx.doi.org/10.1080/13658810902984228.
Guan, Q., Kyriakidis, P.C., Goodchild, M.F., 2011. A parallel computing approach to fast geostatistical areal interpolation. Int. J. Geogr. Inform. Sci. 25 (8), 1241–1267. http://dx.doi.org/10.1080/13658816.2011.563744.
Haklay, M., 2010. How good is volunteered geographical information? A comparative study of openstreetmap and ordnance survey datasets. Environ. Plan. B: Plan. Des. 37 (4), 682–703. http://dx.doi.org/10.1068/b35097.
Harris, R., Singleton, A., Grose, D., Brunsdon, C., Longley, P., 2010. Grid-enabling geographically weighted regression: a case study of participation in higher education in England. Trans. GIS 14 (1), 43–61. http://dx.doi.org/10.1111/j.1467-9671.2009.01181.x.
Haworth, J., Shawe-Taylor, J., Cheng, T., Wang, J., 2014. Local online kernel ridge regression for forecasting of urban travel times. Transport. Res. Part C: Emerging Technol. 46, 151–178. http://dx.doi.org/10.1016/j.trc.2014.05.015.
Hegarty, M., 2011. The cognitive science of visual-spatial displays: implications for design. Topics Cogn. Sci. 3 (3), 446–474. http://dx.doi.org/10.1111/j.1756-8765.2011.01150.
Hegarty, M., Smallman, H.S., Stull, A.T., 2012. Choosing and using geospatial displays: effects of design on performance and metacognition. J. Exp. Psychol.: Appl. 18 (1), 1–17. http://dx.doi.org/10.1037/a0026625.
Heuvelink, G., Griffith, D.A., 2010. Space–time geostatistics for geography: a case study of radiation monitoring across parts of Germany. Geogr. Anal. 42 (2), 161–179. http://dx.doi.org/10.1111/j.1538-4632.2010.00788.x.
Hoffer, D., 2014. What Does Big Data Look Like? Visualisation is Key for Humans. <http://www.wired.com/2014/01/big-data-look-like-visualization-key-humans/> (retrieved May 4, 2015).
Hoffman, J., 2012. Q&A: the data visualizer. Nature 486 (7401). http://dx.doi.org/10.1038/486033a, 33–33.
Hoffmann, M., Raman, R., Muthukrishnan, S., 2007. Streaming algorithms for data in motion. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4614 LNCS, pp. 294–304.
Huang, W., Li, S., Liu, X., Ban, Y., 2015. Predicting human mobility with activity changes. Int. J. Geogr. Inform. Sci. http://dx.doi.org/10.1080/13658816.2015.1033421.
Hughes, J., 1989. Why functional programming matters. Comput. J. 32 (2), 98–107. http://dx.doi.org/10.1093/comjnl/32.2.98.
Jiang, B., 2013. Head/tail breaks: a new classification scheme for data with a heavy-tailed distribution. The Professional Geogr. 65 (3), 482–494. http://dx.doi.org/10.1080/00330124.2012.700499.
Jiang, B., 2015a. Head/tail breaks for visualization of city structure and dynamics. Cities 43, 69–77. http://dx.doi.org/10.1016/j.cities.2014.11.013.
Jiang, B., 2015b. Geospatial analysis requires a different way of thinking: the problem of spatial heterogeneity. GeoJournal 80 (1), 1–13. http://dx.doi.org/10.1007/s10708-014-9537-y.
Jiang, B., Miao, Y., 2015. The evolution of natural cities from the perspective of location-based social media. The Professional Geogr. 67 (2), 295–306. http://dx.doi.org/10.1080/00330124.2014.968886.
Jiang, B., Yin, J., 2014. Ht-index for quantifying the fractal or scaling structure of geographic features. Ann. Assoc. Am. Geogr. 104 (3), 530–540. http://dx.doi.org/10.1080/00045608.2013.834239.
Juditsky, A., Hjalmarsson, H., Benveniste, A., Delyon, B., Ljung, L., Sjöberg, J., Zhang, Q., 1995. Nonlinear black-box models in system identification: mathematical foundations. Automatica 31 (12), 1725–1750. http://dx.doi.org/10.1016/0005-1098(95)00119-1.
Kandwal, R., Augustijn, E.-W., Stein, A., Miscione, G., Garg, P.K., Garg, R.D., 2010. Geospatial analysis of HIV-related social stigma: a study of tested females across Indian mandals. Int. J. Health Geogr. 9 (18).
Kanevski, M., Pozdnoukhov, A., Timonin, V., 2009. Machine Learning for Spatial Environmental Data: Theory, Applications, and Software. EFPL Press.
Kealy, A., Retscher, G., Toth, C., Brzezinska, D., 2014. Collaborative positioning: concepts and approaches for more robust positioning. In: Proceedings of the XXV FIG Congress 2014: Engaging the Challenges Enhancing the Relevance, p. 15.
Keim, D., Qu, H., Ma, K.-L., 2013. Big-data visualization. IEEE Comput. Graphics Appl. 33 (4), 20–21.
Kitchin, R., 2014. Big data, new epistemologies and paradigm shifts. Big Data Soc. 1 (1), 1–12. http://dx.doi.org/10.1177/2053951714528481.
Kitchin, R., 2013. Big data and human geography opportunities, challenges and risks. Dialogues Human Geogr. 3 (3), 262–267. http://dx.doi.org/10.1177/2043820613513388.
Knapp, L., 1995. A task analysis approach to the visualization of geographic data. In: Nyerges, T.L., Mark, D.M., Laurini, R., Egenhofer, M.J. (Eds.), Cognitive Aspects of Human–Computer Interaction for Geographic Information Systems. Springer, Netherlands, pp. 355–371.
Kohli, D., Sliuzas, R., Kerle, N., Stein, A., 2012. An ontology of slums for image-based classification. Comput., Environ. Urban Syst. 36 (2), 154–163. http://dx.doi.org/10.1016/j.compenvurbsys.2011.11.001.
Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., Mostashari, F., 2005. A space-time permutation scan statistic for disease outbreak detection. PLoS Med. 2 (3), e59. http://dx.doi.org/10.1371/journal.pmed.0020059.
Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
Lee, K., Ganti, R.K., Srivatsa, M., Liu, L., 2014. Efficient spatial query processing for big data. In: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '14, pp. 469–472, doi:http://dx.doi.org/10.1145/2666310.2666481.
Li, Y., Yan, D., Liu, S., 2011. Dimensionality reduction algorithm based on density portrayal. Comput. Eng. 37 (21), 138–140.
Liu, Z., Jiang, B., Heer, J., 2013. ImMens: Real-Time Visual Querying of Big Data. In: Computer Graphics Forum (Proceedings of EuroVis '13), 32(3pt4), pp. 421–430.
Longley, P.A., Adnan, M., Lansley, G., 2015. The geotemporal demographics of twitter usage. Environ. Plan. A 47 (2), 465–484. http://dx.doi.org/10.1068/a130122p.
Lu, Y., Zhang, M., Witherspoon, S., Yesha, Y., Yesha, Y., Rishe, N., 2013. sksOpen: efficient indexing, querying, and visualization of geo-spatial big data. In: Proceedings of the 2013 12th International Conference on Machine Learning and Applications (ICMLA), 2, pp. 495–500, doi:http://dx.doi.org/10.1109/ICMLA.2013.196.
Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., Jie, W., 2014. Remote sensing big data computing: challenges and opportunities. Future Generation Comput. Syst. http://dx.doi.org/10.1016/j.future.2014.10.029.
MacEachren, A.M., Crawford, S., Akella, M., Lengerich, G., 2008. Design and implementation of a model, web-based, GIS-enabled cancer atlas. The Cartogr. J. 45 (4), 246–260. http://dx.doi.org/10.1179/174327708X347755.
Maitrey, S., Jha, C.K., 2015. Handling big data efficiently by using map reduce technique. In: Paper presented at the 2015 IEEE International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, IN.
Mandelbrot, B.B., 1982. The Fractal Geometry of Nature. W. H. Freeman, New York.
Mandelbrot, B.B., Hudson, R.L., 2004. The (Mis)behavior of Markets: A Fractal View of Risk, Ruin and Reward. Basic Books, New York.
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011. Big Data: The Next Frontier for Innovation, Competition, and Productivity.
Marshall, W., Boshuizen, C., 2013. Planet Labs' Remote Sensing Satellite System. Small Satellite Conference.
Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt, New York.
McDougall, K., 2011. Using volunteered information to map the Queensland floods. In: Proceedings of the Surveying & Spatial Sciences Biennial Conference 2011, pp. 13–24.
Mechveliani, S.D., 2001. Computer algebra with Haskell: applying functional-categorial-'lazy' programming. In: Proceedings of International Workshop CAAP-2001 (Dubna, Russia), pp. 203–211; <http://ca-d.jinr.ru/confs/CAAP/Final/proceedings/proceed.ps>.
Mennis, J., Liu, J.W., 2005. Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. Trans. GIS 9 (1), 5–17. http://dx.doi.org/10.1111/j.1467-9671.2005.00202.x.
Milewski, B., 2009. Lock options. DR DOBBS J. 34 (1), 28–31.
Miller, H.J., Goodchild, M., 2014. Data-driven geography. GeoJournal (Online First), 1–13.
Miller, H.J., Hanz, J., 2009. Geographic Data Mining and Knowledge Discovery: An Overview Geographic Data Mining and Knowledge Discovery, second ed. CRC Press, pp. 1–26.
Mintchev, S., 2014. User-defined rules made simple with functional programming. In: Abramowicz, W., Kokkinaki, A. (Eds.), Business Information Systems. Springer International Publishing, pp. 229–240.
Mohammed, E.A., Far, B.H., Naugler, C., 2014. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioData Mining 7 (1), 22. http://dx.doi.org/10.1186/1756-0381-7-22.

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

Mondzech, J., Sester, M., 2011. Quality analysis of OpenStreetMap data based on application needs. Cartographica: The Int. J. Geogr. Inform. Geovisualization 46 (2), 115–125.

Montello, D.R., 2002. Cognitive map-design research in the twentieth century: theoretical and empirical approaches. Cartogr. Geogr. Inform. Sci. 29 (3), 283–304. http://dx.doi.org/10.1559/152304002782008503.

Moore, R.E., Kearfott, R.B., Cloud, M.J., 2009. Introduction to Interval Analysis. SIAM, Philadelphia, pp. 223.

Morais, C.D., 2012. Where is the Phrase "80% of Data is Geographic" From?. <http://www.gislounge.com/80-percent-data-is-geographic/> (retrieved May 2, 2015).

Musiige, D., Anton, F., Mioc, D., 2013. RF Subsystem Power Consumption and Induced Radiation Emulation. (Ph.D.), Technical University of Denmark, Kongens Lyngby.

Musiige, D., Anton, F., Yatskevich, V., Vincent, L., Mioc, D., Pierre, N., 2011. RF power consumption emulation optimized with interval valued homotopies. Proc. World Acad. Sci. Eng. Technol. 81, 147–153.

Nakaya, T., Yano, K., 2010. Visualising crime clusters in a space–time cube: an exploratory data-analysis approach using space–time kernel density estimation and scan statistics. Trans. GIS 14 (3), 223–239. http://dx.doi.org/10.1111/j.1467-9671.2010.01194.x.

Neis, P., Zielstra, D., Zipf, A., 2012. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. Future Internet 4 (1), 1–21. http://dx.doi.org/10.3390/fi4010001.

Pettit, C., Stimson, R., Nino-Ruiz, M., Moradini, L., Widjaja, I., Delaney, P., et al., 2014. Supporting Urban Informatics through a Big Data Analytics Online Workbench. NSF Workshop on Big Data and Urban Informatics.

Pfeifer, P.E., Deutsch, S.J., 1980. A three-stage iterative procedure for space–time modelling. Technometrics 22, 35–47.

Poser, K., Kreibich, H., Dransch, D., 2009. Assessing volunteered geographic information for rapid flood damage estimation. In: Proceedings of the 12th AGILE International Conference on Geographic Information Science. Advances in GIScience.

Rajasekar, U., Stein, A., Bijker, W., 2006. Image mining for modeling of forest fires from Meteosat images. IEEE Trans. Geosci. Remote Sensing 45 (1), 246–253.

Richter, K.-F., Winter, S., 2011. Citizens as database: conscious ubiquity in data collection. In: Pfoser, D., Tao, Y., Mouratidis, K., Nascimento, M.A., Mokbel, M., Shekhar, S., Huang, Y. (Eds.), Advances in Spatial and Temporal Databases. Springer, Berlin Heidelberg, pp. 445–448.

Riebeek, H., 2015. Big Data Helps Scientists Dig Deeper. <http://earthobservatory.nasa.gov/Features/LandsatBigData/> (retrieved May 23, 2015).

Roth, R.E., 2013. An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. IEEE Trans. Visual. Comput. Graphics 19 (12), 2356–2365. http://dx.doi.org/10.1109/TVCG.2013.130.

Ruff, J., 2002. Information Overload: Causes, Symptoms and Solutions. Harvard Graduate School of Education's Learning Innovations Laboratory (LILA), pp. 1–13.

Rulinda, C.M., Stein, A., Turdukulov, U.D., 2013. Visualizing and quantifying the movement of vegetative drought using remote-sensing data and GIS. Int. J. Geogr. Inform. Sci. 27 (8), 1481–1496. http://dx.doi.org/10.1080/13658816.2012.723712.

Saha, B., Srivastava, D., 2014. Data quality: the other face of big data. In: Proceedings of the 2014 IEEE 30th International Conference on Data Engineering (ICDE), pp. 1294–1297.

Schnürer, R., Sieber, R., Çöltekin, A., 2014. The next generation of atlas user interfaces: a user study with "digital natives". In: Brus, J., Vondrakova, A., Vozenilek, V. (Eds.), Modern Trends in Cartography. Springer International Publishing, pp. 23–36.

Sester, M., Arsanjani, J.J., Klammer, R., Burghardt, D., Haunert, J.-H., 2014. Integrating and Generalising Volunteered Geographic Information. Abstracting Geographic Information in a Data Rich World – Methodologies and Applications of Map Generalisation. Springer, Heidelberg, pp. 119–155.

Sharifzadeh, M., Shahabi, C., 2009. Approximate voronoi cell computation on spatial data streams. Vldb J. 18 (1), 57–75. http://dx.doi.org/10.1007/s00778-007-0081-y.

Shekhar, S., Evans, M.R., Gunturi, V., Yang, K., Cugler, D.C., 2014. Benchmarking spatial big data. In: Rabl, T., Poess, M., Baru, C., Jacobsen, H.-A. (Eds.), Specifying Big Data Benchmarks. Springer, Berlin Heidelberg, pp. 81–93.

Shekhar, S., Evans, M.R., Kang, J.M., Mohan, P., 2011. Identifying patterns in spatial information: a survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (3), 193–214. http://dx.doi.org/10.1002/widm.25.

Shekhar, S., Gunturi, V., Evans, M.R., Yang, K., 2012. Spatial big-data challenges intersecting mobility and cloud computing. In: Proceedings of the 11th ACM International Workshop on Data Engineering for Wireless and Mobile Access – MobiDE '12. doi:http://dx.doi.org/10.1145/2258056.2258058.

Shneiderman, B., 2014. The big picture for big data: visualization. Science 343 (6172), 730. http://dx.doi.org/10.1126/science.343.6172.730-a.

Sinnott, R.O., Bayliss, C., Bromage, A., Galang, G., Grazioli, G., Greenwood, P., et al., 2015. The Australia urban research gateway. Concurr. Comput.: Practice Exp. 27 (2), 358–375. http://dx.doi.org/10.1002/cpe.3282.

Slocum, T.A., Blok, C., Jiang, B., Koussoulakou, A., Montello, D.R., Fuhrmann, S., Hedley, N.R., 2001. Cognitive and usability issues in geovisualization. Cartogr. Geogr. Inform. Sci. 28 (1), 61–75. http://dx.doi.org/10.1559/152304001782173998.

Slocum, T.A., McMaster, R.B., Kessler, F.C., Hugh, H.H., 2008. Thematic Cartography and Geovisualization, third ed. Prentice Hall.

Steed, C.A., Ricciuto, D.M., Shipman, G., Smith, B., Thornton, P.E., Wang, D., et al., 2013. Big data visual analytics for exploratory earth system simulation analysis. Comput. Geosci. 61, 71–82. http://dx.doi.org/10.1016/j.cageo.2013.07.025.

Stein, A., Groenigen, J.W.v., Jeger, M.J., Hoosbeek, M.R., 1998. Space-time statistics for environmental and agricultural related phenomena. Environ. Ecol. Statistics 5 (2), 155–172.

Straumann, R.K., Çöltekin, A., Andrienko, G., 2014. Towards (re)constructing narratives from georeferenced photographs through visual analytics. The Cartogr. J. 51 (2), 152–165. http://dx.doi.org/10.1179/1743277414Y.0000000079.

Suthaharan, S., 2014. Big data classification: problems and challenges in network intrusion prediction with machine learning. Performance Eval. Rev. 41 (4), 70–73. http://dx.doi.org/10.1145/2627534.2627557.

Tan, H., Luo, W., Ni, L.M., 2012. CloST: A Hadoop-Based Storage System for Big Spatio-Temportal Data Analytics. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2139–2143, doi: http://dx.doi.org/10.1145/2396761.2398589.

Tran, N., Skhiri, S., Lesuisse, A., Zimányi, E., 2012. AROM: Processing big data with Data Flow Graphs and functional programming. In: Paper presented at the 4th IEEE International Conference on Cloud Computing Technology and Science (CloudCom), Taipei.

Turton, I., Openshaw, S., 1998. High-performance computing and geography: developments, issues, and case studies. Environ. Plan. A 30 (10), 1839–1856. http://dx.doi.org/10.1068/a301839.

Ujang, U., Anton, F., Azri, S., Rahman, A.A., Mioc, D., 2014. 3D hilbert space filling curves in 3D city modeling for faster spatial queries. Int. J. 3-D Inform. Model. (IJ3DIM) 3 (2). http://dx.doi.org/10.4018/ij3dim.2014040101.

Umamaheshwaran, R., Bijker, W., Stein, A., 2007. Image mining for modeling of forest fires from meteosat images. IEEE Trans. Geosci. Remote Sensing 45 (1), 246–253. http://dx.doi.org/10.1109/TGRS.2006.883460.

Van de Kassteele, J., Stein, A., 2006. A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output. Environmetrics 17 (4), 309–322. http://dx.doi.org/10.1002/env.771.

Van de Vlag, D., Stein, A., 2006. Modeling dynamic beach objects using spatio-temporal ontologies. J. Environ. Inform. 8 (1), 22–33.

Van de Vlag, D., Vasseur, B., Stein, A., Jeansoulin, R., 2005. An application of problem and product ontologies for the revision of beach nourishments. Int. J. Geogr. Inform. Sci. 19 (10), 1057–1072. http://dx.doi.org/10.1080/13658810500032404.

Van Zyl, T., Simonis, I., McFerren, G., 2009. The sensor web: systems of sensor systems. Int. J. Digital Earth 2 (1), 16–30. http://dx.doi.org/10.1080/17538940802439549.

Veregin, H., 2005. Data quality parameters. In: Goodchild, M.F., Longley, P.A., Maguire, D.J., Rhind, D.W. (Eds.), New Developments in Geographical Information Systems: Principles, Techniques, Management and Applications. Wiley, Hoboken, NY, pp. 177–189.

Vickers, D., Rees, P., 2007. Creating the UK national statistics 2001 output area classification. J. R. Statist. Soc.: Ser. A 170 (2), 379–403. http://dx.doi.org/10.1111/j.1467-985X.2007.00466.x.

Wang, S., Ding, G., Zhong, M., 2013. On spatial data mining under big data. J. China Acad. Electron. Inform. Technol. 8 (1), 8–17.

Wang, S., Yuan, H., 2013. Spatial data mining in the context of big data. In: Paper presented at the The 2013 International Conference on Parallel and Distributed Systems (ICPADS), Seoul, KW.

Wang, S., Yuan, H., 2014. Spatial data mining: a perspective of big data. Int. J. Data Warehousing Mining 10 (4), 50–70. http://dx.doi.org/10.4018/ijdwm.2014100103.

Wood, S.A., Guerry, A.D., Silver, J.M., Lacayo, M., 2013. Using social media to quantify nature-based tourism and recreation. Sci. Reports 3, 2976. http://dx.doi.org/10.1038/srep02976.

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., et al., 2012. Visual analytics for the big data era – a comparative review of state-of-the-art commercial systems. In: Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182, doi:http://dx.doi.org/10.1109/VAST.2012.6400554.

Zielstra, D., Zipf, A., 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. In: Proceedings of the 13th AGILE International Conference on Geographic Information Science 2010.

Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley Press, Oxford, UK.

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