International Journal of
Geo-Information

Editorial

Introduction to Big Data Computing for Geospatial Applications

Zhenlong Li [1,*], Wenwu Tang [2], Qunying Huang [3], Eric Shook [4] and Qingfeng Guan [5]

[1] Geoinformation and Big Data Research Laboratory, Department of Geography, University of South Carolina, Columbia, SC 29208, USA

[2] Center for Applied Geographic Information Science, Department of Geography and Earth Sciences, University of North Carolina at Charlotte, Charlotte, NC 28223, USA; wenwutang@uncc.edu

[3] Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA; qhuang46@wisc.edu

[4] Department of Geography, Environment, and Society, University of Minnesota, Minneapolis, MN 55455, USA; eshook@umn.edu

[5] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China; guanqf@cug.edu.cn

[*] Correspondence: zhenlong@sc.edu

Received: 3 August 2020; Accepted: 10 August 2020; Published: 12 August 2020

Abstract: The convergence of big data and geospatial computing has brought challenges and opportunities to GIScience with regards to geospatial data management, processing, analysis, modeling, and visualization. This special issue highlights recent advancements in integrating new computing approaches, spatial methods, and data management strategies to tackle geospatial big data challenges and meanwhile demonstrates the opportunities for using big data for geospatial applications. Crucial to the advancements highlighted here is the integration of computational thinking and spatial thinking and the transformation of abstract ideas and models to concrete data structures and algorithms. This editorial first introduces the background and motivation of this special issue followed by an overview of the ten included articles. Conclusion and future research directions are provided in the last section.

Keywords: geospatial big data; geospatial computing; cyberGIS; GeoAI; spatial thinking

1. Introduction

Earth observation systems and model simulations are generating massive volumes of disparate, dynamic, and geographically distributed geospatial data with increasingly finer spatiotemporal resolutions [1]. Meanwhile, the ubiquity of smart devices, location-based sensors, and social media platforms provide extensive geo-information about daily life activities. Efficiently analyzing those geospatial big data streams enables us to investigate complex patterns and develop new decision-support systems, thus providing unprecedented values for sciences, engineering, and business. However, handling the five "Vs" (volume, variety, velocity, veracity, and value) of geospatial big data is a challenging task as they often need to be processed, analyzed, and visualized in the context of dynamic space and time [2].

Following a series of successful sessions organized at the American Association of Geographers (AAG) Annual Meeting since 2015, this special issue on "Big Data Computing for Geospatial Applications" by the *ISPRS International Journal of Geo-Information* aims to capture the latest efforts on utilizing, adapting, and developing new computing approaches, spatial methods, and data management strategies to tackle geospatial big data challenges for supporting applications in different domains, such as climate change, disaster management, human dynamics, public health, and environment and engineering.

Specifically, this special issue aims to address the following important topics: (1) geo-cyberinfrastructure integrating spatiotemporal principles and advanced computational technologies (e.g., GPU (graphics processing unit computing), multicore computing, high-performance computing, and cloud computing); (2) innovations in developing computing and programming frameworks and architecture (e.g., MapReduce, Spark) or parallel computing algorithms for geospatial applications; (3) new geospatial data management strategies and storage models coupled with high-performance computing for efficient data query, retrieval, and processing (e.g., new spatiotemporal indexing mechanisms); (4) new computing methods considering spatiotemporal collocation (locations and relationships) of users, data, and computing resources; (5) geospatial big data processing, mining, and visualization methods using high-performance computing and artificial intelligence; (6) integrating scientific workflows in cloud computing and/or a high-performance computing environment; and (7) other research, development, education, and visions related to geospatial big data computing. This editorial provides a summary of the ten articles included in this issue and suggests future research directions in this area based on our collective observations.

2. Overview of the Articles

The articles included in this issue make significant contributions to the use of big data computing for tackling various geospatial problems (from human mobility to disaster management to knowledge discovery) by incorporating novel methodologies, data structures, and algorithms with advanced computing frameworks (from geo-visual analytics, deep learning to cloud computing, and MapReduce/Spark). Using ten different big data sources (e.g., social media, remote sensing, and Internet of Things), this issue demonstrates the value and importance of integrating computational approaches and geospatial methods in advancing scientific discovery and domain applications (Table 1).

Table 1. Summary of the geospatial big data and computing approaches used in each article for various geospatial applications.

Category	Geospatial Application	Big Data Source	Computing Approaches	Article
Big Data Computational Methods	Geospatial data preprocessing	Sensor data via Internet of Things (IoT)	Parallel extracting, transforming, loading, MapReduce/Hadoop	Jo and Lee. (2019) [3]
	Overlay analysis	Land use (as a case study)	High performance computing with Spark, cloud computing	Zhao et al. (2019) [4]
	Land-use change prediction	Remote sensing (Landsat)	Parallel modeling with MapReduce/Hadoop, cloud computing	Kang et al. (2019) [5]
	Global scale terrain analysis	Global elevation	Google Earth Engine, cloud computing	Safanelli et al. (2020) [6]
Big Data Mining	Human mobility (pattern discovery)	Public transit	Machine learning (clustering algorithm), visual analytics	Zhang et al. (2019) [7]
	Disaster management (earthquake mitigation)	Social media	Deep learning (CNN), spatiotemporal analysis	Yang et al. (2019) [8]
	Missing road generation	Navigation (trajectory)	A set of new computing algorithms	Wu et al. (2019) [9]
Knowledge Representation	Geospatial problem solving	Heterogeneous data via online services	Workflow, online geoprocessing, knowledge base	Zhuang et al. (2018) [10]
	Geographic knowledge representation	Ontological	Knowledge graph, ontologies	Wang et al. (2019) [11]
Big Data Search	Geospatial big data management and searching (climate data)	Climate	Cyberinfrastructure-based cataloging, spatiotemporal indexing	Gaigalas et al. (2019) [12]

2.1. Big Data Computational Methods

Geospatial data processing and analysis, such as transformation in geometry, converting coordination reference systems, and evaluating spatial relationships, often include a large number of floating-point arithmetic computations. Correspondingly, MapReduce and Spark-based frameworks and systems, such as SpatialHadoop [13] and GeoSpark [14], were developed to speed up these computations. Additionally, cloud-based computing platforms, such as Google Earth Engine (GEE) for big earth observation data, have been increasingly used in geospatial studies and applications. To optimize the performance of a parallel algorithm for geospatial processing, analysis, or modeling when using such general-purpose frameworks, the spatial characteristics of the data and algorithm must be considered for the algorithmic design [15,16]. The four papers by Jo et al. [3], Zhao et al. [4], Kang et al. [5], and Safanelli et al. [6] focus on parallel computing and highlight the adaption of existing computing frameworks for geospatial data preprocessing, parallel algorithm design, simulation modeling, and data analysis.

It often takes a long time to prepare geospatial datasets for these data computing systems, which generally involves extracting, transforming, and loading (i.e., ETL) processes. To deal with big data in the ETL process, Jo and Lee proposed a new method, D_ELT (delayed extracting–loading–transforming), to reduce the time required for data transformation within the Hadoop platform by utilizing MapReduce-based parallelization [3]. Using big sensor data of various sizes and geospatial analysis of varying complexity levels, several experiments are performed to measure the overall performance of D_ELT, traditional ETL, and extracting–loading–transforming (ELT) systems. Their results demonstrate that D_ELT outperforms both ETL and ELT. In addition, the larger the amount of data or the higher the complexity of the analysis, the better the performance of D_ELT over the traditional ETL and ELT approaches.

Zhao et al. designed a parallel algorithm for overlay analysis, which uses a measurement of polygon shape complexity as the key factor for data partitioning in combination with a distributed spatial index and a minimum boundary rectangular filter [4]. The parallel algorithm was implemented based on Spark, a widely used distributed computing framework for large-scale applications [17]. Experiment results show data partitioning based on shape complexity effectively improved the load balancing among multiple computing nodes, hence the computational efficiency of the parallel algorithm. This work demonstrates that appropriate definitions and measurements of the properties of data and/or algorithms (no matter how simple they are) to reflect the computational intensities are of essential significance for the performance enhancement of parallel algorithms.

The CA–Markov model is one of the most widely used extended cellular automata (CA) models and has been used in the prediction and simulation of land-use changes [5]. As land-use change simulation and prediction involves massive amounts of data and calculations, many parallel CA algorithms have been designed to simulate urban growth based on various computing models, including central processing units (CPUs) and GPUs. While the parallel CA method incorporates spatial relationships amongst cells, it cannot maintain connections between partitions after a study area is divided into several pieces, resulting in different prediction results. Meanwhile, the traditional Markov method can maintain integrity for the entire study area but lacks the ability to incorporate spatial relationships amongst the cells. Alternatively, the MapReduce framework is capable of efficient parallel processing when coupled to the CA–Markov model; the key problem of segmentation and maintaining spatial connections remain unresolved. As such, Kang et al. introduced a MapReduce-based solution to improve the parallel CA–Markov model for land-use-change prediction [5]. Results suggest that the parallel CA–Markov model not only solves the paradox that the traditional CA–Markov model cannot simultaneously achieve the integrity and segmentation for land-use change simulation and prediction but also achieves both efficiency and accuracy.

Safanelli et al. took a different approach to handle geospatial big data challenges [6]. They developed a terrain analysis algorithm based on GEE (termed TAGEE) to calculate a variety of terrain attributes, e.g., slope, aspect, and curvatures, for different resolutions and geographical extents.

By using spheroidal geometries measured by the great-circle distance, TAGEE does not require the input DEM data to be projected on a flat plane. Experiments show that TAGEE can generate similar results when compared to conventional GIS software packages. By taking advantage of the high-performance computing capacity of GEE, TAGEE is able to efficiently produce a suite of terrain attribute products at any spatial resolution at a global scale. This work represents an emerging paradigm of geospatial computing in the era of big data. As cloud computing platforms such as GEE mature, geospatial computing is no longer limited by locally available computing resources and datasets. Applications of complex geospatial algorithms/models at high spatial resolutions and the global scale have been seen in the last couple of years and will soon become the norm.

2.2. Big Data Mining

Social sensing, in which humans represent a large sensor network, has emerged as a new data collection approach in the big data era [18]. The following three papers by Zhang et al. [7], Yang et al. [8], and Wu et al. [9] demonstrate the power of integrating social sensing data (public transit, social media, and mobile phone) and big data computing techniques for supporting geospatial applications including human mobility, disaster management, and transportation.

Zhang et al. developed a novel approach for mining and visualizing human mobility patterns from multisource big public transit data, aiming to support transportation planning and management by providing an enhanced understanding of human movement patterns over space and time [7]. To efficiently extract travel patterns from massive heterogeneous data sources, this work developed a clustering algorithm to extract transit corridors indicating the connections between different regions and a graph-embedding algorithm to reveal hierarchical mobility community structures. Beyond the novel machine-learning algorithms, this work also provides a scalable web-based geo-visual analytical system including visualization techniques to allow users to interactively explore the extracted patterns. The system was evaluated by 23 users with different backgrounds and the results confirm the usability and efficiency of the integrated geo-visual analytical approach for human movement pattern discovery from public transit big data. This work demonstrates the power of integrating geospatial big data, machine-learning algorithms, and geo-visual analytical approaches for supporting transportation applications.

Yang et al. introduced a deep learning method to efficiently conduct sentiment analysis of big social media data for assisting disaster mitigation [8]. This work devises a five-phase framework for automatic extraction of public emotions from geotagged Sina micro-blog data including data collection and processing, emotion classification, and spatiotemporal analysis. To classify emotion (fearful, anxious, sad, angry, neutral, and positive), a convolutional neural network (CNN) model is designed and trained by converting the raw text to a word vector. To demonstrate the efficiency of the approach, an earthquake in Ya'an, China, in 2013 was used as a case study. Based on the trained model, public emotions within the study area are classified at different time periods right after the earthquake. Spatiotemporal analyses were then performed to examine the dynamics of people's sentiments toward the earthquake over space and time. Results suggest that the proposed approach accurately classified emotions from big social media data (>81%), providing valuable public emotional information for disaster mitigation.

Wu et al. proposed a three-step approach to detect missing road segments from mobile phone-based navigation data within urban environments [9]. Their first step is to apply filtering to navigation data to remove those related to pedestrian movement and existing road segments. Then, as a second step, centerlines of missing roads are constructed using a clustering algorithm. Building the topology of missing roads and connecting these detected roads with existing road networks is the third step. Wu et al. [9] applied this approach in a study area (about 6 square kilometers) in Shanghai, China. Based on ~10 million GPS points collected from mobile navigation in 2017, this work evaluated the capability of their three-step approach in the detection of missing roads. Results demonstrate the performance of this three-step approach based on mobile phone data, recognizing the computational challenge of their approach when dealing with larger datasets.

2.3. Knowledge Representation

Zhuang et al. [10] addressed an understudied problem, namely the representation and sharing of knowledge related to geospatial problem solving. Through a process of abstraction and decomposition, this work deconstructs geospatial problems into tasks that operate at three different granularities. Beyond a high-level description, this work formalizes the geospatial problem-solving process into a knowledge base by creating a suite of ontologies for tasks, processes, and GIS operations. Using a meteorological early-warning analysis as a case study, this work successfully demonstrates how to capture abstract geospatial problem-solving knowledge in a formal and sharable task-oriented knowledge base. Demonstrated by a prototype system, their results offer a promising glimpse of how users could begin building geospatial problem-solving models and workflows similar to spatial models and workflows. Such models and workflows could be re-used and adapted for similar problems or used as a building block to tackle more complex geospatial problems in the future, such as the global effects caused by climate change.

Wang et al. [11] built a knowledge graph similar to Zhuang et al. [16] but instead focused on capturing geographic objects and their spatiotemporal contexts. This work creates a geographic knowledge graph (GeoKG) comprised of six elements to answer foundational questions in geography including: Where is it? Why is it there? When and how did it happen? Through a process of model construction and formalization, this work captures geographic objects, their relations, and ongoing dynamics in a GeoKG. To demonstrate the effectiveness of the GeoKG, this work detailed the evolution of administrative divisions of Nanjing, China, along the Yangzi River and then compared it to a well-known straightforward and extensible ontology known as YAGO (Yet Another Great Ontology). Results show that GeoKG improved accuracy and completeness through analyses and user evaluation, demonstrating scientific advancement in capturing geographic knowledge in a computational system.

2.4. Big Data Search

Lastly, Gaigalas et al. [12] presented a cyberinfrastructure-enabled cataloging approach that combines web services and crawler technologies to support efficient search of big climate data. The cataloging approach consists of four main steps, including selection and analysis of a metadata repository, crawling of metadata using crawlers, building spatiotemporal indexing of metadata, and search based on collection search (via catalog services) and granule search (via REST API). This cataloging approach was implemented to support EarthCube CyberConnector. To demonstrate the feasibility and efficiency of the proposed approach, this cyberinfrastructure was tested with petabyte-level ESOM (Earth System Observation and Modeling) data provided by UCAR THREDDS Data Server (TDS). Results suggest that the proposed cataloging approach not only boosts the crawling speed by 10 times but also dramatically reduces the redundant metadata from 1.85 gigabytes to 2.2 megabytes. Instead of focusing on big data analysis, this work demonstrates the significance and advanced techniques of making big climate data searchable to support interdisciplinary collaboration in climate analysis.

3. Conclusion and Future Research Directions

This special issue highlights a diversity of geospatial models and analyses, geospatial data, geospatial thinking, and computational thinking used to address myriad geospatial problems ranging from human mobility [7] to disaster management [8]. The manuscripts span geospatial problem solving and knowledge (e.g., [10,11]), handling massive geospatial data (e.g., [3,12]), and analyzing and visualizing geospatial data (e.g., [7,9]).

Crucial to the advancements highlighted in this special issue is the integration of computational thinking and spatial thinking and the translation of abstract ideas and models to concrete data structures and algorithms. A promising future research direction will be to build on this integration of knowledge and skills across the disciplines of GIScience and computational science, which has been termed cyber literacy for GIScience [19]. In this way, integrated knowledge of real-world geospatial patterns and computational processes can be captured and shared, and big data and geospatial visual analytic frameworks can be integrated to provide more robust computational geospatial platforms to address myriad geospatial problems. A key challenge in this research direction will be the integrative fabric that can seamlessly combine scholarly thinking with computational infrastructures, geospatial data elements with big data capabilities, and geospatial methods infused with parallelism.

Parallelism can be achieved by innovatively utilizing advanced computing frameworks, such as MapReduce and Spark, for applications that include massive data sorting, computing, machine learning, and graph processing [20]. While this special issue highlighted advancements in geospatial big data preprocessing [3], land-use change prediction [5], and overlay analysis [4], more efforts should be devoted to identifying geospatial applications of great impact and benefiting from the integration of geospatial methods and parallelization in the big data era. Additionally, many existing geospatial big data applications simply inject spatial data types or functions inside existing big data systems (e.g., Hadoop) without much optimization [3]. Therefore, further research directions should focus on improving and optimizing the performance of big data frameworks from different aspects, such as data ETL, job scheduling, resource allocation, query analytics, memory issues, and I/O bottlenecks, by considering the spatial principles and constraints [21].

Author Contributions: Conceptualization, Zhenlong Li; Writing—original draft preparation, Zhenlong Li, Wenwu Tang, Qunying Huang, Eric Shook and Qingfeng Guan; writing—review and editing, Zhenlong Li, Eric Shook, Wenwu Tang, Qunying Huang and Qingfeng Guan. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We would like to thank all authors for contributing to this special issue and the reviewers for their constructive suggestions and criticisms that significantly improved the papers in this issue. We also want to thank Yuanyuan Yang for her editorial support and help in preparing the special issue.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, Z.; Yang, C.; Jin, B.; Yu, M.; Liu, K.; Sun, M.; Zhan, M. Enabling big geoscience data analytics with a cloud-based, MapReduce-enabled and service-oriented workflow framework. *PLoS ONE* **2015**, *10*, e0116781. [CrossRef] [PubMed]
2. Li, Z. Geospatial Big Data Handling with High Performance Computing: Current Approaches and Future Directions. In *High Performance Computing for Geospatial Applications*; Tang, W., Wang, S., Eds.; Springer: New York, NY, USA, 2020; ISBN 978-3-030-47997-8.
3. Jo, J.; Lee, K.-W. Map Reduce-Based D_ELT Framework to Address the Challenges of Geospatial Big Data. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 475. [CrossRef]
4. Zhao, K.; Jin, B.; Fan, H.; Song, W.; Zhou, S.; Jiang, Y. High-Performance Overlay Analysis of Massive Geographic Polygons That Considers Shape Complexity in a Cloud Environment. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 290. [CrossRef]
5. Kang, J.; Fang, L.; Li, S.; Wang, X. Parallel Cellular Automata Markov Model for Land Use Change Prediction over MapReduce Framework. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 454. [CrossRef]
6. Safanelli, J.L.; Poppiel, R.R.; Ruiz, L.F.C.; Bonfatti, B.R.; Mello, F.A.d.O.; Rizzo, R.; Demattê, J.A.M. Terrain Analysis in Google Earth Engine: A Method Adapted for High-Performance Global-Scale Analysis. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 400. [CrossRef]
7. Zhang, T.; Wang, J.; Cui, C.; Li, Y.; He, W.; Lu, Y.; Qiao, Q. Integrating Geovisual Analytics with Machine Learning for Human Mobility Pattern Discovery. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 434. [CrossRef]
8. Yang, T.; Xie, J.; Li, G.; Mou, N.; Li, Z.; Tian, C.; Zhao, J. Social Media Big Data Mining and Spatio-Temporal Analysis on Public Emotions for Disaster Mitigation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 29. [CrossRef]

9. Wu, H.; Xu, Z.; Wu, G. A Novel Method of Missing Road Generation in City Blocks Based on Big Mobile Navigation Trajectory Data. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 142. [CrossRef]
10. Zhuang, C.; Xie, Z.; Ma, K.; Guo, M.; Wu, L. A Task-Oriented Knowledge Base for Geospatial Problem-Solving. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 423. [CrossRef]
11. Wang, S.; Zhang, X.; Ye, P.; Du, M.; Lu, Y.; Xue, H. Geographic Knowledge Graph (GeoKG): A Formalized Geographic Knowledge Representation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 184. [CrossRef]
12. Gaigalas, J.; Di, L.; Sun, Z. Advanced Cyberinfrastructure to Enable Search of Big Climate Datasets in THREDDS. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 494. [CrossRef]
13. Eldawy, A. SpatialHadoop: Towards flexible and scalable spatial processing using MapReduce. In Proceedings of the SIGMOD Ph.D. Symposium 2014, Snowbird, UT, USA, 22 June 2014; pp. 46–50.
14. Yu, J.; Wu, J.; Sarwat, M. Geospark: A cluster computing framework for processing large-scale spatial data. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Bellevue, WA, USA, 3–6 November 2015; p. 70.
15. Guan, Q.; Zeng, W.; Gong, J.; Yun, S. pRPL 2.0: Improving the parallel raster processing library. *Trans. GIS* **2014**, *18*, 25–52. [CrossRef]
16. Li, Z.; Hodgson, M.E.; Li, W. A general-purpose framework for parallel processing of large-scale LiDAR data. *Int. J. Digit. Earth* **2018**, *11*, 26–47. [CrossRef]
17. Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Ghodsi, A. Apache Spark: A unified engine for big data processing. *Commun. ACM* **2016**, *59*, 56–65. [CrossRef]
18. Li, Z.; Huang, Q.; Emrich, C. Introduction to Social Sensing and Big Data Computing for Disaster Management. *Int. J. Digit. Earth* **2019**, *12*, 1198–1204. [CrossRef]
19. Shook, E.; Bowlick, F.J.; Kemp, K.K.; Ahlqvist, O.; Carbajeles-Dale, P.; Di Biase, D.; Rush, J. Cyber literacy for GIScience: Toward formalizing geospatial computing education. *Prof. Geogr.* **2019**, *71*, 221–238. [CrossRef]
20. Li, Z.; Hu, F.; Schnase, J.L.; Duffy, D.Q.; Lee, T.; Bowen, M.K.; Yang, C. A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 17–35. [CrossRef]
21. Yang, C.; Wu, H.; Huang, Q.; Li, Z.; Li, J. Using spatial principles to optimize distributed computing for enabling the physical science discoveries. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 5498–5503. [CrossRef] [PubMed]