Chapter 2
A Short Note on Data-Intensive Geospatial Computing

Bin Jiang

For over 1,000 years, human knowledge has been recorded in printed formats such as books, journals, and reports archived and stored in libraries or museums. In the past decades, ever since the invention of computers, human knowledge can also be hidden in digital data. Although part of the data has been researched and reported, the pace of the data explosion has been dramatic and exceeds what the printed medium can deal with. This is particularly true for the twenty-first century, enormous data volume of scientific data, including geospatial data (NRC 2003), collected by all kinds of instruments in a variety of disciplines mostly on a 24/7 basis. The lack of related cyberinfrastructure for the data capture, curation and analysis, and for communication and publication alike led Jim Gray to lay out his vision of the fourth research paradigm – data-intensive science (Bell et al. 2009). This fourth paradigm differs fundamentally, in terms of the techniques and technologies involved, from the third paradigm of computational science (ca. 50 years ago) on simulating complex phenomena or processes. Before the computer age, there was only empirical science (ca. 1,000 years ago), and then theoretical science (ca. 500 years ago) like Newton's Laws of Motion and Maxwell's Equations. Nowadays, scientists do not look through telescopes but instead mine massive data for research and scientific discoveries. Every discipline x has evolved, over the past decades, into two computer- or information-oriented branches, namely x-informatics and computational x (Gray 2007, cited from Hey et al. 2009). Geography is no exception. Both geoinformatics, which deals with the data collection and analysis of geoinformation, and computational geography or geocomputation (Gahegan 1999), which has to do with simulating geographic phenomena and processes, are concurrent research issues under the banner of geographic information science.

The data size or scale under the fourth paradigm and for data-intensive computing is up to terabytes or petabytes using high-performance supercomputers.

B. Jiang
Department of Technology and Built Environment, Division of Geomatics, University of Gävle, 80176 Gavle, Sweden
e-mail: bin.jiang@hig.se

V. Popovich et al. (eds.), *Information Fusion and Geographic Information Systems: Towards the Digital Ocean*, Lecture Notes in Geoinformation and Cartography 5,
DOI 10.1007/978-3-642-19766-6_2,

Table 2.1 Evolution of PC in terms of cost and data size [the last two columns are partially according to Steve Kopp's talk (Kopp 2010)]

Year	Data size (byte)	1 GB hard disk ($)	256 MB RAM ($)
1980	1	223,000	200,000
1990	1,000	10,000	50,000
2000	1,000,000	10	50
2010	1,000,000,000	0.1	5

What data scale do we adopt daily and using an ordinary personal computer (PC) for geospatial analysis then? Let us assume that we were able to deal with data scale in bytes in the 1980s. According to Moore's law, up to this year 2010 we should be able to deal with gigabytes (Table 2.1). Ask ourselves, what data scale are we dealing with in analysis and computation today? Most of us are probably still constrained at the scale of megabytes if not kilobytes. It seems we are far behind the time, using far less computing power of the modern PC. In the mean time, if we take a look at the cost of 1 GB storage and 256 MB memory of a PC, it has been plummeting over the past three decades since 1981 (Table 2.1), when IBM introduced a PC powered by an Intel 8088 processor. In what follows, I will use three examples to illustrate that the scale of gigabytes is quite feasible for the state-of-the-art PC or affordable server-like machines.

Triggered by an obvious morphological difference between European cities (being organic or self-organized) and US cities (being more planned in nature), I started a study trying to illustrate the difference in a quantitative manner. This was the original study motivation. However, instead of the seemingly obvious difference, the study ended up with a similarity – a universal pattern of urban street networks (Jiang 2007). This pattern can be described by the 80/20 principle: about 80% of streets are less connected (below an average, e.g., 4 or 5), whereas 20% of streets are highly connected (above the average). This discovery was quite a surprise to me. In the course of the investigation, 40 cities with different sizes (10 largest cities, 10 smallest cities, and 20 middle-sized cities) were chosen from over 4,000 US cities, and all bear this pattern or regularity. This pattern holds true even for the smallest US town Duffield in Virginia. This study goes beyond US cities and includes some cities elsewhere from Israel, China, and Germany. The pattern perceived still remains valid. All these data have been archived in the web for free access, http://fromto.hig.se/~bjg/PhyAData/. In the study, the total size of the data examined is up to a couple of gigabytes. The data we used in the study is the freely available US Census Bureau TIGER database.

GPS data constitute one of the emerging geospatial data formats over the past decade or so. Pervasive GPS units with various mobile devices provide an unprecedented way of collecting individual-based geospatial data. Similar individual-based data include the data from mobile phone transmitters about phone users' daily movement trajectories. In one of the studies with the aim to explore GIS-based mobility information for sustainable urban planning, we collaborated with a local taxi company and collected massive GPS data (over 6 GB) about human movement patterns. We analyzed over 72,000 people's moving trajectories, obtained from

50 taxicabs during a 6-month period in a large street network involving four cities. The human movement demonstrates a scaling in terms of travel length (see video clips http://fromto.hig.se/~bjg/gps/). We illustrated that goal-oriented moving behavior has little effect on the overall traffic distribution (Jiang et al. 2009; see applet available at http://fromto.hig.se/~bjg/movingbehavior/). We further implemented some agent-based simulations (with both random and purposive moving agents) which took days in a state-of-the-art personal computer to get one result, in total a couple of weeks computer time to get all the simulations done. The study ended up with a finding that purposive, or random moving behavior, has little effect on the overall mobility pattern (Jiang and Jia 2011). This finding indicates that given a street network, the movement patterns generated by purposive human beings and by random walkers have little difference. We are still bound by a non-disclosure agreement, for the taxi company is reluctant to release the part of the GPS data for research purposes. However, we have made the massive simulated data available at (http://fromto.hig.se/~bjg/PREData/).

Not only analysis but also computation can be done at a gigabytes scale for geospatial data. For example, the computation of shortest paths tends to be intensive and time consuming, in particular when a graph is huge, involving hundreds of millions of vertices and edges. Because of the enormity of road networks, and the request of real-time computation of shortest paths in map services like Google Maps, it would need dozens of gigabytes of memory to accommodate the enormous road network of the entire world. In this connection, the only data source available at this scale, which meets the request of testing is probably OpenStreetMap (OSM), a wiki-like collaboration to create free editable map of the world using data from portable GPS units, aerial photography, and other free sources. Since its inception in year 2004, it has now attracted over 200,000 of registered users (Jackson 2010). OpenStreetMap constitutes one of the most successful examples of crowdsourcing – massive amateurs performing functions that were used to be performed by trained professionals, or volunteered geographic information (VGI) – a term coined by Goodchild (2007) to refer to the user-generated geographic information on the Web. Using the OSM data, we have been developing an intelligent routing service – FromToMap (http://www.fromtomap.org/). We have conducted some data-intensive computation for deriving the fewest-turn routes. For Europe, there are over 30 GB OSM data, from which we extracted 17 GB street data. For the purpose of computing routes, we generated a huge graph involving 10 million nodes and 17 million links, occupying about 30 GB of memory. In addition, we implemented algorithms to compute isochrones within a few minutes far from any location. The FromToMap service is hosted by an affordable HP server, costing about $7,000 with 48 GB memory and a 1 TB hard disk.

From the above elaboration, let me add a few reflections on data-intensive geospatial computing. First of all, geospatial research should go beyond the data scale of kilobytes and megabytes. If raw data are in gigabytes, there is probably no need to sample the data. Geographic theory and knowledge should be based on some massive data for verification, to avoid the distorted picture of the elephant in the eyes of the blind man. Current computing storage and processing

capacity are fairly powerful in dealing with the data scale. Second, geographic space is heterogeneously distributed. The larger the geographic space, the more heterogeneous it appears. For example, some physicists have extracted massive trajectory data of registered dollar notes (another format of VGI) from http://www.wheresgeorge.com/ to study human mobility patterns that bear this heterogeneity (Brockmann et al. 2006). In this regard, I believe that geographers or geospatial researchers have unique contributions to the research using the cutting edge geographic information technologies. Third, OSM can be a benchmark data for geospatial research. OSM data is very rich in content. It contains streets, pedestrian, and cycling paths, as well as public transports. In addition, points of interest and land use information are also embedded in the data. More importantly, there is no constraint to get access to it, since it is owned by no one. This point seems applicable to other formats of VGI as well.

Following the third point, geospatial community should set up a data repository to archive data, algorithms, and source codes that can be shared among interested researchers. This way geospatial research is based on some common datasets, becoming replicable, extendible, and reusable in terms of algorithms and codes in line with open-source activities. Similar efforts have been attempted in many other disciplines such as biology, physics, and computer sciences. Finally, many analytics tools or geovisualization tools we have developed do not meet the challenge of data-intensive computing. And many tools developed are still targeted to verify a given hypothesis and to find some presumed relationship. Current tools are still rather weak in discovering hidden knowledge. Computing power and physical memory limits for operation systems to operate have been increasing dramatically. Virtually, there is no upper limit for the recently released Windows 2008. In the meantime, cloud computing has fully deployed for those who need more computing power than ordinary PC. In this regard, many Internet giants such as Amazon, eBay, Google, and Microsoft all have their own cloud computing services (Armbrust et al. 2009). GIS software suppliers like ESRI have also been working on the next generation of products that are to be more cloud compatible. We are facing an exciting yet challenging age when geospatial analysis and computation need to be done at a massive scale.

Acknowledgments An early version of this chapter was presented at the National Science Foundation TeraGrid Workshop on Cyber-GIS, February 2–3, 2010, Washington DC.

References

Armbrust M., Fox A., Griffith R., Joseph A. D., Katz R. H., Konwinski A., Lee G., Patterson D. A., Rabkin A., Stoica I., and Zaharia M. (2009), Above the Clouds: A Berkeley View of Cloud Computing, Technical Report available at http://www.eecs.berkeley.edu/Pubs/TechRpts/

Bell G., Hey T. and Szalay A. (2009), Beyond the data deluge, *Science*, 323, 1297–8.

Brockmann D., Hufnage L., and Geisel T. (2006), The scaling laws of human travel, *Nature*, 439, 462–465.

Gahegan M. (1999), What is geocomputation? *Transactions in GIS*, 3(3), 203–206.
Goodchild M. F. (2007), Citizens as sensors: the world of volunteered geography, *GeoJournal*, 69 (4), 211–221.
Hey T., Tansley S. and Tolle K. (2009), *The Fourth Paradigm: data intensive scientific discovery*, Microsoft Research, Redmond, Washington.
Jackson J. (2010), OpenStreetMap Attracts 200,000 Volunteers, PCWorld, http://www.pcworld.com/article/186096/openstreetmap_attracts_200000_volunteers.html
Jiang B. (2007), A topological pattern of urban street networks: universality and peculiarity, *Physica A: Statistical Mechanics and its Applications*, 384, 647–655.
Jiang B. and Jia T. (2011), Agent-based simulation of human movement shaped by the underlying street structure, *International Journal of Geographical Information Science*, 25, 51–64, http://arxiv.org/abs/0910.3055.
Jiang B., Yin J. and Zhao S. (2009), Characterizing human mobility patterns in a large street network, *Physical Review E*, 80, 021136, Preprint, arXiv:0809.5001.
Kopp S. (2010), Toward spatial modeling in the cloud, A position paper presented at the National Science Foundation TeraGrid Workshop on Cyber-GIS, February 2–3, 2010 – Washington DC.
NRC (2003), *IT Roadmap to a Geospatial Future*, The National Academies Press: Washington, D. C.

http://www.springer.com/978-3-642-19765-9

Information Fusion and Geographic Information Systems
Towards the Digital Ocean
Popovich, V.; Claramunt, C.; Devogele, Th.; Schrenk, M.; Korolenko, K. (Eds.)
2011, XII, 180 p., Hardcover
ISBN: 978-3-642-19765-9