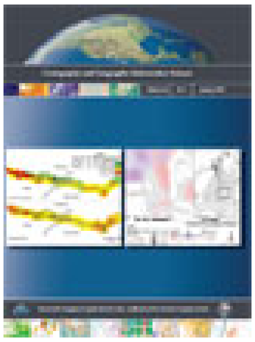


Cartography and Geographic Information Science

ISSN: 1523-0406 (Print) 1545-0465 (Online) Journal homepage: http://www.tandfonline.com/loi/tcag20

The Internet of Things and fast data streams: prospects for geospatial data science in emerging information ecosystems

Marc P. Armstrong, Shaowen Wang & Zhe Zhang

To cite this article: Marc P. Armstrong, Shaowen Wang & Zhe Zhang (2018): The Internet of Things and fast data streams: prospects for geospatial data science in emerging information ecosystems, Cartography and Geographic Information Science, DOI: 10.1080/15230406.2018.1503973

To link to this article: https://doi.org/10.1080/15230406.2018.1503973

Published online: 13 Sep 2018.

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE
https://doi.org/10.1080/15230406.2018.1503973

Taylor & Francis
Taylor & Francis Group

The Internet of Things and fast data streams: prospects for geospatial data science in emerging information ecosystems

Marc P. Armstrong [a], Shaowen Wang [b] and Zhe Zhang [b]

[a]Department of Geographical and Sustainability Sciences, The University of Iowa, IA, USA; [b]Department of Geography and Geographic Information Science, University of Illinois, Urbana, IL, USA

ABSTRACT

This paper surveys the rapid development of the Internet of Things, the massive data streams that are only now beginning to be generated from it, and the resulting opportunities and challenges that these data streams bring to geographic information analysis. These challenges arise because streaming data volumes cannot be subjected to analysis using the standard repertoire of methods that have been designed to analyze static geospatial datasets. New approaches are needed, not to supplant, but to supplement, these existing tools. A focus is placed on the concept of data velocity (fast data) and its effects on sampling and inference. Innovative data ingestion strategies based on principles related to reservoir sampling and sketching are described. Dynamic temporal data flows present significant challenges to load balancing in distributed (e.g. cloud) parallel environments, even at exascale levels of performance. Further advances in the exploitation of data locality based on geographical concepts, as well as advanced processing methods based on edge and approximate computing, require further elucidation. Concepts are illustrated using a database compiled from a distributed sensor network of mobile radioactivity detectors.

ARTICLE HISTORY
Received 16 May 2018
Accepted 20 July 2018

KEYWORDS
Internet of Things; sensors; data streams; fast data; parallel computing

Introduction

We are now in the midst of a wave of transformative change in the rate of production of geospatial data. The purpose of this paper is describe these advances and to address the challenges that are encountered when geospatial methods are applied to these emerging data sources. A particular focus is placed on rapidly streaming geospatial information generated by the Internet of Things and its effect on geospatial sampling strategies and analysis.

A quick look backward will help to illustrate the magnitude of the changes that have occurred. The "First Symposium on Geographical Information Systems" took place in Ottawa, Ontario in late September, 1970. The Symposium, sponsored by the International Geographical Union Commission on Geographical Data Sensing and Processing, drew 49 participants. The Summary of Proceedings from that gathering comments on the quantity of "pieces" of data (defined as recognizable elements of real world data) that are associated with each identifier, ranging from the first category which consists of one (a bit plane) to the sixth and largest category (Tomlinson, 1970, p. 10, emphasis added) that "contains a *very large number of pieces of data* associated with each location identifier." In this case, the very large number means that there are more than 5000 attributes (presumably an integer, real or character string) per identifier. It is telling that no system of that era was able to handle such miniscule, static data volumes.

Clearly, the science and technology of geographic information handling have advanced in ways that were effectively unimaginable in 1970. And yet we have now arrived at the cusp of a new paradigm of data acquisition and analysis (NRC, 2013). This paradigm is characterized by the collection and transmission of massive, streaming data volumes, which gives rise to the need to develop highly efficient analysis methods. One term that is applied to these massive data quantities is "big data." The National Institute of Science and Technology (NIST, 2015, p. 4) parenthetically defines some of the terms that are often used:

> "*Big Data refers to the inability of traditional data architectures to efficiently handle the new datasets. Characteristics of Big Data that force new architectures are* ***volume*** *(i.e., the size of the dataset) and* ***variety*** *(i.e., data from multiple repositories, domains, or types), and the data in motion characteristics of* ***velocity*** *(i.e.,*

CONTACT Marc P. Armstrong marc-armstrong@uiowa.edu

> *rate of flow) and* ***variability*** *(i.e., the change in other characteristics).*"

Volume and variety have been a concern of geographers and geographic information science for decades. Berry (1964), for example, defined a geographic matrix that would accommodate an enormous variety of data types over a wide range of spatial and temporal scales. Haggett and Chorley (1967, p. 30) address the volume issue directly when, in discussing Berry's paper, they refer to "the explosion of the data matrix" and many other authors have remarked on the rapid pace of data developments (for roughly decadal benchmarks, see Armstrong, 2000; Beaumont, 1989; Calkins, 1990; Miller, 2010). During the past few years, research specifically addressing the concept of geospatial big data has emerged (e.g., Li et al., 2016; Shekhar, Evans, Gunturi, Yang, & Cugler, 2014; Wiener et al., 2016; Yang, Yu, Jiang, & Li, 2017). It is only recently, however, that velocity has emerged as a dimension of interest because most geographic information in the past could best be characterized as "data at rest." For example, the decennial US Census takes a snapshot of socio-economic conditions every ten years. Snapshots, of course, can be animated much as a flip-book consists of a sequence of static images. But big data streams far exceed that which has been experienced heretofore (see, e.g., Shekhar et al., 2014).

Several long-standing research areas overlap partially with the analysis of streaming geospatial data, yet each is somehow deficient in the current context. Snapshots of data at rest can be situated in the research domain of temporal GIS. Early work that sketched out initial concepts (Armstrong, 1988; Langran, 1988; Langran & Chrisman, 1988) has given way to subsequent research that elucidated further challenges and provided suggestions for substantive improvements (Langran, 1992; Peuquet, 1994; Peuquet & Duan, 1995; Wachowicz, 1999; Worboys, 1994; Yuan & Stewart, 2007). Other researchers have studied humans as they move through time and space. Hägerstrand (1970) pioneered this "time geography" approach, which is summarized by Thrift (1977). Several of the main themes of Hägerstrand's work are placed in a modern GIS research context and have been extended by, for example, Miller (1991, 2005), Kwan (1998), and Shaw (2006). While these analyses are significant, they also tended to be retrospective, rather than real-time in construction.

Additional areas of geographic research impinge on the analysis of big data streams. Geographic data mining has traditionally been focused on the extraction of information and knowledge from large static databases and data warehouses, but as noted by Miller and Han (2009), this pattern has changed as demonstrated in papers by, for example, Yuan (2009), Cao, Mamoulis, and Cheung (2009), and Laube and Duckham (2009). This change is enabled by the rapid evolution in the capabilities of mobile devices and wireless sensors.

In short, geospatial research has long been concerned about the general concept of big data (and its temporal aspects), even though its definition continues to evolve. However, the technological milieu has shifted the data acquisition ground toward decentralized devices that are streaming massive amounts of data. Current geospatial analysis methods, which have their roots in the previous century, are unable to cope with this emerging flood. New approaches are required to help data scientists gain insight from these increasingly prominent sources of information.

The emerging information infrastructure of the Internet of Things

The widespread availability of mobile computing and inexpensive sensors with radios has created a new data ecosystem, the Internet of Things (IoT), which refers to the incorporation of digital components into a vast array of "things," both large and small. This change has now placed data *velocity* in the spotlight: real-time geographic information that is streamed from sensors has brought us into an era of fast data. Consider an example reported by Jarr (2015, p. 20) that consists of 53 million electric meters streaming usage information several times each second in order to monitor changes in demand and provide feedback to "smart" systems about variable charges that can, in turn, have an environmental impact by holding down demand peaks.

In addition to being in widespread use, the IoT has become institutionalized in different contexts, not the least of which is the US Senate in the form of Senate Resolution 110 of the 1st Session of the 114th Congress (http://thomas.loc.gov/cgi-bin/query/z?c114:S.RES.110.ATS:/).As further evidence of its broad acceptance, the IoT now plays an endorsed role in the operations of a certain large, well-known software firm (headquartered in the Pacific Northwest) which has built IoT support into the latest version of its operating system, where it is referred to as "IoT Core." This software is designed specifically to run on lightweight devices with a minimum overhead footprint.

This increase in the deployment of devices and sensors has been continuously fueled by rapid, steady declines in the cost of electronic components. These decreases have taken place in accordance with a "law" (Denning & Lewis, 2017; Moore, 1965) as related to the density of transistors on an integrated circuit, and are paralleled by improvements in storage capacity and network speed, among other factors. WiFi speeds now often enter into the 1 Gbs range, meaning that sensors are able to stream high resolution data at very high sampling rates. While WiFi and cellular coverage is clearly not universally available, problems with communication coverage gaps are being filled by large satellite constellations. For example, Eutelsat's geostationary satellite constellation provides Ka-band broadband services to European and North American markets (de Selding, 2015) and has recently expanded coverage into14 sub-Saharan nations. Other companies are either coming online or are planning services (OneWeb, LeoSat, SpaceX) using large constellations (100s) of low-earth-orbit satellites (Henry, 2018). Capabilities provided by such constellations support global sensor network communication needs. While these trends may smack of Whigish technological determinism, these changes are undeniable and persistent.

Though advances in communication and sensing are its key enablers, the IoT is about much more than connectivity. Porter and Heppelmann (2014, p. 66) assert that we are on the brink of a third wave of competition and innovation that is driven by information technologies. The first, was the transformation from analog to digital. The second transformation took place as a consequence of Internet connectivity which supported integration of business functions and supply chains. In each of these two cases, however, objects themselves were untransformed. The IoT changes that with embedded sensors, computers, and network links fundamentally changing the nature of products and what they can do. The following sections briefly outline changes occurring in three areas with geographical implications: vehicle telematics, distributed sensor networks, and advances in remote sensing.

Vehicle telematics

New automobiles have a multitude of data acquisition systems that feed real-time data processing capabilities. Sensors monitor vehicle dynamics and include various types of streaming remote sensing systems that operate in visible (cameras and LiDAR), infrared (night vision), and radar (vehicle sensing) frequencies. This information is transported via a VANET (vehicular ad hoc network) (Lochert, Scheuermann, & Mauve, 2010; Wang, 2018). Functions typically supported by telematics systems include routing, enhanced visualization, lane departure and automated parking. Systems that are under development enable autonomous vehicles to communicate with other vehicles (and the transportation infrastructure) in order to reduce congestion and accident likelihood; these functions also depend on the acquisition and real-time processing of massive amounts of data (see, e.g., Wang & Wets, 2013). At present there is a wide difference of opinion about the practical feasibility of autonomous vehicles operation (Calo, 2018; Eckhoff & Sommer, 2014; GAO (Government Accountability Office), 2013; Kirkpatrick, 2015).

Sensor networks

Applications of sensor networks are growing at a very rapid pace. Paralleling other electronic devices, GNSS (global navigation satellite system) receivers have decreased rapidly in size, cost, and power consumption. This technology is critical for monitoring a wide variety of sensor inputs, and a variety of small form-factor and ultra-low-power products provide high-accuracy locations and often include barometer, hygrometer, and compass options (Malkos, del Castillo and Mole, 2104). Many sensors are locationally fixed and comprise an essential element of "smart city" applications. There is considerable interest by large technology companies, including AT&T and Microsoft, in this arena. In some cases, urban sensors monitor particulates (Rajasegarar et al., 2014) and temperatures (Mohring, Myers, Atkinson, VanDerWal, & van der Valk, 2015; Mone, 2015, p. 20) *in situ*. In other cases, fixed sensors are positioned in nonurban environments to achieve a sampling objective. One example is the Jefferson Project at Lake George, NY that monitors, literally, streaming watershed characteristics (Romero, 2015). In yet other cases, the sensors are mobile, locationally aware, and transported by vehicles or worn by people (Gupta, Holloway, Heravi, & Hailes, 2015; Ilyas, Alwakeel, Alwakeel, & Aggoune, 2014). Though most sensor networks are concerned primarily with monitoring, others close the loop with feedback control over some system and perform actions on themselves or the environment (Nayak & Stojmenovic, 2010; Serpanos, 2018).

Environmental remote sensing

The resolution of remote sensing technology continues to increase across spatial, spectral, radiometric, and temporal dimensions. Two trends are important: LiDAR and new satellite constellations that provide streaming video data.

LiDAR information is now widely available for display and analysis. Current LiDAR systems collect massive amounts of data in the form of point clouds, with some systems capable of generating > 200,000 3D coordinates each second. As a consequence of this data volume, researchers are turning to the use of high performance computers to extract information from the generated point clouds (Hegeman, Sardeshmukh, Sugumaran, & Armstrong, 2014). Satellite remote sensing is also experiencing a major transformation. Monolithic systems, such as Landsat, are being supplemented by new form factors such as cubesats with a 10 × 10 × 10 cm nominal unit size (1U) representing one liter and an approximate weight of one kilogram. These small-form-factor satellites are being placed into orbit by several private-sector entities though ownership of these constellations is in a state of flux. For example, Skybox became Terra Bella, which was acquired by Google in 2014, and was later sold to Planet Labs, which had its own constellation of SkySat satellites. The original Skybox configuration uses a mini-fridge satellite form factor (weight is approximately 100 kg) to acquire submeter imagery that can be streamed as video (Butler, 2014); each one generates more than a terabyte of data each day in the form of sub-meter images and 1.1 m resolution video streaming at a rate of 30 frames each second. Planet now has more than 175 satellites providing medium and high resolution images (SkySat has a 0.8 m resolution with subweekly revisit capabilities). It is not clear whether these ventures are sustainable given current market demand for imagery. In addition, there are important questions about the density of the proposed constellations and the possibility of collision cascades that could turn orbits into junkyards. Foust (2015) reports that some disparage these orbiting platforms as “debris sats” despite international and US government guidelines that require satellites to be deorbited after 25 years (Honda, Perkins, & Sun, 2013).

In sum, the increasingly inexpensive devices with shrinking form factors that comprise the IoT have a variety of sensor types, communicate wirelessly, and are locatable. This evolving information ecosystem is generating massive amounts of streamed sensor data that has important implications for the way that geospatial data are managed and analyzed.

The effect of data volume and velocity on research paradigms

Connected streaming devices are contributing to what is now called the fourth paradigm of scientific research (Hey, Tansley, & Tolle, 2009). The four paradigms, however, are not crisply demarcated, and research activities often flow among them. Since the first three paradigms are already well-defined in the geographical literature, in this section, a particular focus is placed on the fourth paradigm.

Paradigm one: observational and experimental

Scientific perspectives evolved independently in multiple places, but advanced most coherently in Greece (Weinberg, 2015). Early scientific practices were based on empirical observations of natural phenomena. Later, controlled experiments based on the scientific method were designed to evaluate hypotheses and show how experimental inputs affected outputs. This approach remains central to current era scientific advances.

Paradigm two: theory and models

Theoretical science has its roots in the work of Newton in the 1600s, though some of Newton’s work was based on the precise measurements (Paradigm One) of planetary motion made by Tycho Brahe. Experiments are usually designed to confirm or falsify theoretical models and there is often a feedback loop between theory and observational practice. Theoretical models have played an important role in geographic investigations for at least half a century (Chorley & Haggett, 1967). Models consistently, but imperfectly, explain the world around us, which gave rise to a quote attributed to the statistician George Box: “All models are wrong, but some are useful” (c.f., Box, 1976, p. 792). He was simply observing that models are simplified abstractions and that any such abstraction will contain errors (Lowry, 1968).

Paradigm three: simulation and computation

Computation, a key enabler of scientific progress, now complements empiricism and theory. An important focus of computational methods in geography has been placed on the development of simulation models, which, in geography were enabled by the increased availability of general purpose digital computers in the late 1960s (e.g. Hägerstrand, 1967; Marble, 2015; Marble & Anderson, 1972). Simulation and other computational approaches to science cohered into what is now called computational science with a direct extension to geographical problems that is sometimes referred to as computational geography (Armstrong, 2000; Openshaw, 1998; Torrens, 2010).

Paradigm four: data-intensive discovery

The Moore's Law-like advances in data collection noted earlier have led to the creation of data streams that exceed our ability to validate, analyze, visualize, and curate them. Indeed much of the data generated by current systems is not intended for human evaluation. Rather, it is data of, by and for machines, and its volume is reportedly more than doubling every eighteen months (NIST, 2015, p. 4). Given the increased volume and pace of data acquisition, a new interdisciplinary data science is emerging. In geography, this view was anticipated more than two decades ago by Openshaw (1995) who advanced an argument for data-driven analysis, a perspective that has been significantly reinforced since then. Data-driven discovery permits a different way of envisioning the relationship between theory and observation, effectively turning it on its head, for if we ask what pattern of observations yields a desired effect, we can discover new relationships that can then lead to theoretical insight. It is important, however, not to forget about the directional effects of correlation and causation, and also to remain mindful that a Popperian view of falsification can play an important role in strengthening or demolishing discovered relations. This is where predictive power becomes important, for if a relation can be shown to be a robust predictor, where robustness refers to persistence, it gains credence and becomes worthy of further investigation and theorizing (Dhar, 2013).

Dhar (2013, p. 70) goes on to describe the need for fluidity in model creation during data-driven exploration, and suggests that, particularly for nonstationary problems, models are but rough predictive approximations that can be periodically adjusted to compensate for changes observed in data streams. This perspective is most relevant in application domains that are characterized less by highly deterministic physical processes, and more by those that must incorporate behavioral variability, such as much of human geography. A failure to make appropriate adjustments in such cases will lead to errors that tend to come from three main sources:

(1) Misspecification. A simple example is fitting a linear model to a nonlinear process, which would introduce bias.
(2) Sampling. Sample size affects parameter estimation; small samples may be biased.
(3) Randomness. Random errors are prominent in most models of human behavior.

Large volumes of data and abundant computational power can reduce the effects of the first two types of errors. Small samples become a thing of the past and model fitting can be done in a large number of ways to improve predictive performance. With large numbers of observations, the effects of random errors also may be reduced considerably. Indeed, Dhar (2013, p. 72) asserts that social scientists have never had a better ability to observe human behaviors and that massive quantities of data "makes induction not only feasible but productive." The view is particularly promising because there is a distinct adaptive approach to analysis that undergirds it. This adaptive computation paradigm is designed explicitly to work in novel, open-ended, dynamic environments (Forrest & Mitchell, 2016).

Induction can be usefully applied to fast data streams to gain insights into processes. While different definitions can be found, Holland, Holyoak, Nisbett, and Thagard (1986, p. 1) suggest that induction encompasses all inferential processes that expand knowledge in the face of uncertainty. During an inductive analysis of a data stream, evidence in the form of new observations is accumulated in order to arrive at a conclusion. Inductive methodologies may best be thought of as truth estimation techniques designed to yield the best possible answer given the information available. The process of inductive pattern evaluation can be made operational using a classifier system (Holland et al., 1986) to process the data stream. Classifiers, often implemented using bit-map pattern matching, can be applied to temporal and spatial series (Bennett & Armstrong, 1989, 1996) and since they are relatively lightweight processes, and can operate in parallel, they hold promise in data stream analysis. This approach is consistent with the use of inductive tools to classify geospatial information in other contexts as argued by Gahegan (2000a, 2000b, 2003)).

Clearly, however, inductive inference is not foolproof; as we move from the specific to the general it is possible to draw incorrect conclusions. An illustrative syllogism often involves (birds, feathers, flight, and penguins). It then follows that induction is context dependent and must have some feedback mechanism in place so that current knowledge can be corrected or enhanced. In addition, the correctness or robustness of an induced relation can be tested by evaluating predictions made about future states of the system under investigation. Predictions are important pursuits in scientific inquiry and may be made on the basis of formal theories arrived at through deduction, or based on inductive processes. If our theories and models are correct, we should be able to make predictions about future system states. If the prediction proves to be incorrect, that will require a revision to the theory or model that gave rise to the prediction.

With the ability to specify inductive models, it has become possible to generate interesting questions, ones that a human might not have considered. These questions can then provide the basis for models that can be stress-tested by means of their ability to make predictions about new information gleaned from a data stream.

Implications of data velocity for geographic information analysis

When current methods of spatial analysis are used to analyze static data sources, and when such data conform to expectations about distributional characteristics, traditional frequentist methods may be applied. In other cases, Bayesian methods are sometimes adopted. Such approaches, however, suffer from very high levels of computational complexity (e.g. Yan, Cowles, Wang, & Armstrong, 2007). New methods of spatial analysis are required to preprocess and analyze large quantities of fast data. Such data have unknown stationarity and error signatures, and as a result spatial data stream methods must be able to accommodate variability in streamed inputs. This poses enormous problems to analysis.

Figure 1 shows how data streams are ingested and immediately subjected to fast space-time analytics (FaST) that require real-time (or near-real-time) response. Such computations are performed during the actual time that an external process occurs, in order that the computation results can be used to control, monitor, or respond in a timely manner to the external process (IEEE, 1990, p. 61). This requires high performance computing as well as new analytical methods that have lightweight computational complexity. The observations are then exported to a big data store where traditional methods of analysis can be applied. An examination of a simplified data value chain can help to develop this conceptualization a bit more (Figure 2).

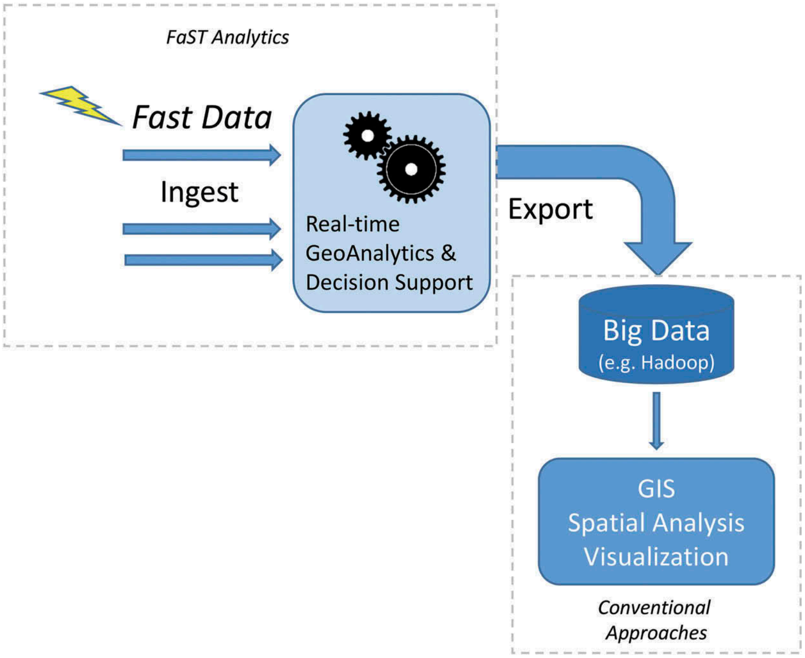


Figure 1. Fast space-time analytics (FaST) operate on fast data immediately upon ingestion and pass them to a data repository for conventional analyses (modified from Jarr, 2015).

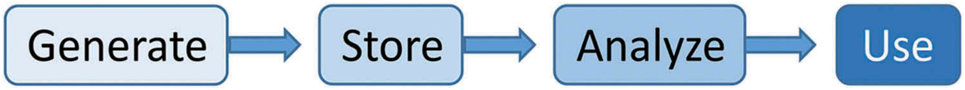


Figure 2. A simple data value chain.

A conventional chain would be formed, for example, when observations about households are submitted as a part of the decennial census. In such cases, observations are sent to a centralized store, where the results are aggregated and analyzed, and then sent back to users who can use the processed information. It is here where data currency becomes important. A reimagined data value chain differentiates each data element according to its "age" and also whether it is a singleton value or has been aggregated (Figure 3). Though fresh data has its greatest value when it is an individual item, its value as an individual item declines over time as it is replaced by new observations in the data stream. However, observations regain value and strength as they age and are then aggregated with other values to yield additional synoptic insights.

An individual geospatial observation normally will not, by itself, wield much analytical power. In the business world, however, a single transaction (e.g. a stock "sell" order) or data item (from a click-stream) can be important, and as a consequence, there is often a high premium placed on real-time response. In contrast, most spatial analysis methods have not pursued a goal of real-time response (cf. Xiong & Marble, 1996). In fact, some GIS-based analyses and spatial modeling applications may require hours to produce results, thus reducing their effectiveness in decision support contexts, and motivating research conducted to improve the performance of compute-intensive methods of geospatial analysis.

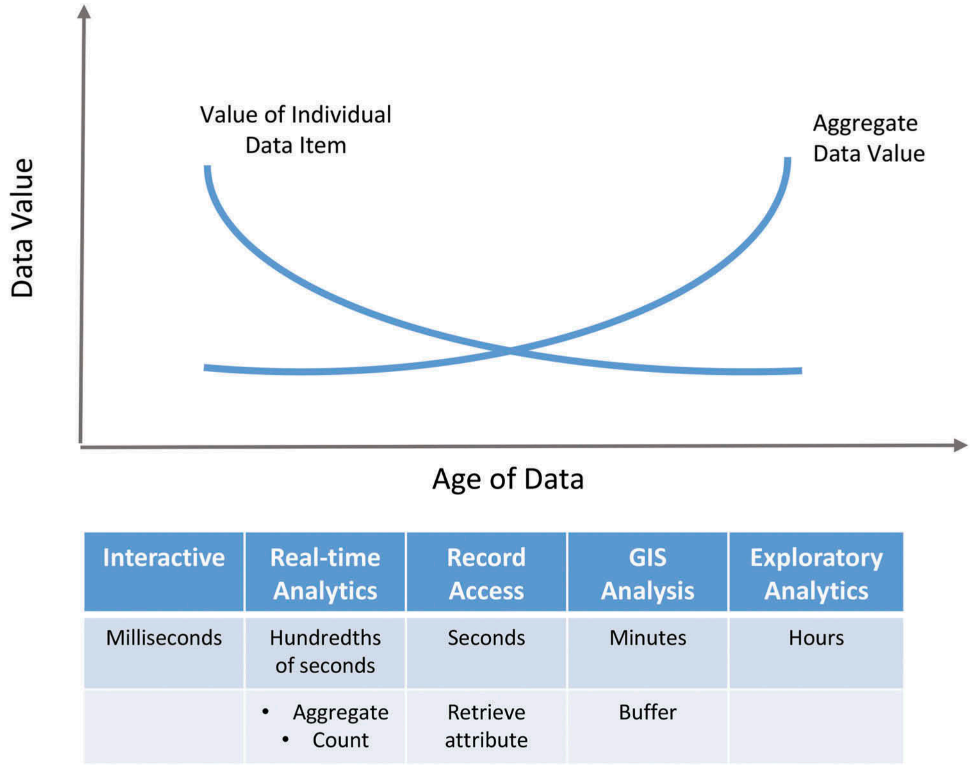
Computing services are now providing low level tools that can be used to construct application-specific solutions to geospatial data streaming problems. Google, for example, has released its Cloud Data Flow service for streaming data on the Google Compute Engine. In a similar vein, Apache Kafka has been designed as a platform that enables users to handle real-time streams (Narkhede, Shapira, & Palino, 2017). These general purpose tools can be used to provide front-end (ingestion) support for the development of geospatial applications.

Streaming analytical task types

With streaming data comes new perspectives on what is possible to contemplate with spatial analytical tools. The four main types of capabilities supported by the IoT in business and industrial applications (Porter & Heppelmann, 2014) have parallels in spatial analysis:

Monitoring

These are the simplest and least costly applications, with sensors measuring and reporting on



Interactive	Real-time Analytics	Record Access	GIS Analysis	Exploratory Analytics
Milliseconds	Hundredths of seconds	Seconds	Minutes	Hours
	• Aggregate • Count	Retrieve attribute	Buffer	

Figure 3. Changes in the form of a data value chain that incorporates fast data streams (modified from Jarr, 2015).

environmental conditions and system performance. The simplest problem is one of detection. However, streams may contain nonzero values below detection limits, and must be carefully evaluated (Helsel, 2005). In other cases, a measured value is reported. For example, a sensor might report on the temperature (and location) of a refrigerated cargo unit. Methods as simple as setting a simple threshold value (e.g. 0°C), would involve one logical operation for each data stream element.

Control

Information obtained by sensors may also be used to effectuate change in some characteristic of a system. For example, a detector designed to monitor toxic aerosols could trigger an alarm long before a human would sense a problem. In the geographical realm, a "geofence" defines a zone that will trigger an action if a mobile device enters (or leaves) it. Other more complex methods continuously evaluate streams to determine when a shift in stream values has occurred in the temporal or spatial domains and form a feedback loop to change inputs in a way analogous to that of a thermostat. Another approach, derived from the field of process control determines whether a process is *in control* or *out of control* (Järpe, 2002; Rogerson, 2009) and triggers an alarm when an out of control state is reached (e.g. a shift in the mean). This is an area of active research in health (syndromic surveillance) and crime analysis (e.g. Eck, Chainey, Cameron, Leitner, & Wilson, 2005; Mandl et al., 2004; Struchen, Vial, & Andersson, 2017).

Optimization

Geographical optimization problems are computationally complex and heuristic methods are sometimes used to address such problems and to support solution space exploration (e.g. Bennett, Xiao, & Armstrong, 2004). The feedback loop that exists between monitoring and control systems in the previous section can be changed to minimize or maximize one or more objectives, such as reduced cost. And smart city applications monitor the human and physical environment in an attempt to optimize urban systems (see, e.g. Deakin, 2013; Vilajosana et al., 2013). Optimization of data streams is a tricky business, however, since any optimal solution would likely get pushed into a suboptimal state with the arrival of new observations.

Autonomy. At this highest level, a system performs all three lower-level functions and interacts with other aspects of the environment to operate independently and either halts or provides notification to a responsible human if the system moves outside some predefined performance envelope.

In each of these cases, the age of each data element is important, as is system response time in the face of massive data volumes. These characteristics raise significant challenges that remain to be addressed in the design and implementation of geospatial methods that can be used to analyze streaming data.

Approaches to improvement of geospatial analytical practices

The following sections are meant to provide an overview of selected areas that may be useful for further exploration as fast data applications are developed. The topics fall into two general categories that are meant to be illustrative rather than exhaustive: (1) data-reduction strategies that do not introduce bias and (2) algorithmic approaches that are intended to reduce the amount of computation needed to produce a useful result.

Data-reduction strategies

Reservoir sampling

Evaluation and analysis of stream data is a one-pass operation that by definition cannot be global, since the full extent of the stream is unknowable. As a consequence, normal sampling procedures cannot be used. In some cases, load shedding, discarding observations randomly or by design, can be used to reduce processing requirements, though this can introduce bias. Reservoir sampling (Vitter, 1985) may be applied to data streams to reduce the magnitude of load shedding bias. In this approach, the first n stream elements comprise the original reservoir, though some implementations sample randomly to get an initial pool. As subsequent elements are streamed each becomes a candidate for inclusion in the reservoir based on an evaluation of its location (temporal index) in the stream. Vitter's approach cuts the evaluation phase by calculating how many records can be skipped before another should be evaluated for inclusion in the reservoir. A modified approach increases the likelihood that more recently added observations are retained in the reservoir sample (Ellis, 2014, p. 327); new data has greater salience than older observations.

Sketching

Sketching has been developed to create synopses and reduce the dimensionality of streaming data. For each new streamed element, a sketch can be developed to determine set membership (has this element appeared before or is it a member of a predefined set?), cardinality (how many different types have appeared in the stream?) and frequency. Put another way, if we have a set ***S*** it would be useful to be able to add additional elements to it, to test whether a new element is already a member of the set, and if true, increment a counter. Ellis (2014, p. 331) states that sketch algorithms have three desirable features:

(1) Data updates are performed in constant time;
(2) Storage space is independent of stream size; and
(3) Queries are performed in linear time for the worst case.

A Bloom filter is one widely adopted approach to sketching (Bloom, 1970) that enables the efficient determination of "heavy hitters" or most frequent values in the stream (Cormode, Korn, Muthukrishnan, & Srivastava, 2008). Useful surveys of sketching approaches are provided by Aggarwal and Yu (2007) and Cormode, Garofalakis, Haas, and Jermaine (2012).

Incremental Clustering. Other approaches to stream characterization are based on incremental clustering. As described by Silva et al. (2013), ancillary data about clusters (microclusters or cluster features) can be stored in the form of a triple CF = (*N, S, SS*), where:

N is the number of data points;

S is a vector that stores the sum of the *N* points; and

SS is a vector that stores the square sum of the *N* points.

Microclusters are incremental; for microcluster *A*, if a streamed element *x* is added to it:

$$\begin{aligned} S_A &\leftarrow S_A + x \\ SS_A &\leftarrow SS_A + x^2 \\ N_A &\leftarrow N_A + 1 \end{aligned}$$

This incrementality allows subclusters to be merged.

$$\begin{aligned} S_C &\leftarrow S_A + S_B \\ SS_C &\leftarrow SS_A + SS_B \\ N_C &\leftarrow N_A + N_B \end{aligned}$$

Aggarwal, Han, Wang, and Yu (2007) describe an algorithm that uses a modified *k*-means approach in which each new element is evaluated (based on root mean square distance deviation) with respect to existing clusters to determine membership in an existing closest cluster or whether a new cluster should be formed.

Progressive analysis

An approach that is focused on latency reduction and is similar to incremental clustering is progressive analysis. The approach has been developed to analyze data sets that are unbounded, providing partial or approximate results at different points during the input stream. Progressive analyses, therefore, are able to provide results that comply with the cognitive constraints of interactive computing, which is usually set at under ten seconds (Fekete & Primet, 2016, p. 1; see also Turkay, Kaya, Balcisoy, & Hauser, 2017).

Computation reduction

Approximate computing

Approximate computing is increasing in importance as a means to improve energy efficiency: shortcuts reduce energy-consuming cycles. The premise is that some applications do not require absolute correctness (Kugler, 2015). This is not a particularly new idea, as "lossy" algorithms are used to encode pictures and music, and location is a variable that is often subjected to approximation (e.g. aggregation). According to Moreau, Sampson, and Ceze (Moreau, Sampson, & Ceze, 2015, p. 12) approximate computing is particularly relevant in mobile environments that involve sensor data collection and summarization.

Sublinear Time Algorithms

Algorithms that execute in linear time represent a "holy grail" of efficiency, though polynomial time is usually considered acceptable. However, if the goal of spatial analysis is to move from low expectations about real time response, a different view will have to be adopted, when data streams on the order of terabytes per second are encountered. Sublinear time algorithms that make assumptions about data distributions to yield answers that are imprecise. While this may be anathema to some, Rubinfeld and Shapira (Rubinfeld & Shapira, 2011, p. 1562) suggest that "there are many situations in which a fast approximate solution is more useful than a slower exact solution." Geographic theory about expected values and locations may prove useful in this regard. Tobler's First Law seems particularly useful here, as does central place theory.

Implementation considerations

Considerable challenges must be overcome in order to assimilate fast data streams and subject them to analyses with real-time responses. Cloud computing has been touted as a solution to such computational needs, as the approach is elastic and designed to be responsive to changing demands (NAS, 2016, p. 111). Geographic researchers have begun to investigate the use of cloud computing in several application domains such as remote sensing (Hegeman et al., 2014; Yang, Xu, &

Nebert, 2013). Yet with enormous numbers of connected devices, each generating massive quantities of data, cloud services may become inefficient, particularly if latency concerns are important. As a consequence, researchers are turning to alternative relationships between devices and the cloud, advocating that processing be pushed from the cloud to the sensor array (fog or edge computing). In the edge paradigm, spatially distributed sensors produce data streams that are preprocessed locally before being transferred to the cloud for more substantial computation and archival storage. Researchers have reported decreases in both latency and power consumption with the edge approach (Shi & Dustdar, 2016).

Many of the sensors that comprise the IoT are limited in their ability to sense and process. daCosta (2013) describes a three-tiered system in which end nodes, akin to a leaf, pass data ("chirps") to propagator nodes that use limited built-in processing to pass the data to integrators that perform higher levels of analysis and control. Chirps are lightweight, disposable, noncritical (possibly redundant), and do not require the end sensor to run the software necessary to support the IP communication stack. Collected chirps are passed to propagator nodes that prune (eliminate redundancy), bundle, summarize, and provide additional context information, such as location. Despite this progress, there are several open issues that require additional research (Shi & Dustdar, 2016, p. 80; Bertino, Nepal, & Ranjan, 2015):

- *Sensor Deployment and Management*: Challenges include placement techniques, duty cycle management and error monitoring.
- *Provenance*: Information about the source of information is important to evaluating its quality.
- *Programmability*: Concerns what data gets processed on the edge before being sent to the cloud. Choices have implications for scalability, though programming frameworks have yet to be developed.
- *Naming*: Edge nodes must be identified. One approach is to use an IP address, though other naming conventions that more robustly support mobile topology reconfiguration may prove superior.
- *Privacy and Security*: Cloud services and data transfers travel over the Internet and are subject to security breaches. Distributed devices also pose security risks since they can be hacked on the edge (Agarwal & Dey, 2016; Noor, Sheng, Maamar, & Zeadally, 2016).

Ingestion

Though spatial data stream methods must be optimized for high rates of ingestion, I/O performance remains a generic problem for computer systems because of fundamental architectural issues (VonNeumann bottleneck). With the current shift toward in-memory analyses and databases, this problem may become diminished though it will never be eliminated (Hegeman et al., 2014; Zhang, Chen, Ooi, Tan, & Zhang, 2015). Even when parallelism is used to increase performance, Amdahl's law (program speedup is limited by the time required to execute its sequential fraction) may rear its ugly head (Amdahl, 1967). Other I/O problems arise when data are streamed from multiple sources, possibly with different provenances and sampling characteristics (e.g. quantization levels). For example, incoming data streams will remain at a constant rate when time-based sampling is used (e.g. 0.1 s). In other cases, event driven sampling may occur and steam rate ingestion will possibly be highly variable.

Processing requirements

Performance is critical when processing data streams. If analytical methods execute more slowly than the incoming data stream, latency will be an increasing function of time. This leads to a consideration of alternative architectural arrangements. If analyses do not require data communication among sensors, then each sensor could be assigned its own processor, a naturally coarse-grained approach (see Figure 4). With spatial data analytics, however, it is likely that local (neighborhood-based) processing will be required and this will introduce communication overhead. In cases where sensors are fixed in place, establishing a neighborhood is relatively straightforward. When they are mobile, the establishment of dynamic neighborhoods becomes problematic. One approach is to define a deformable triangulation with sensor locations forming mobile vertices. There are many triangulation algorithms that tend to fall into general families. The serial algorithm that seems to be most promising is the incremental insertion approach (Tsai, 1993), though it still has a time complexity of $\boldsymbol{O}$ (n log n). Wu, Guan, and Gong (2011) and Hegeman et al. (2014) report on parallel implementations, but computational performance remains an open issue. Nevertheless, parallelism should hold a prominent place in future work on streaming spatial data, for it is clear that US science policy will continue to press forward with high-performance (parallel) computing as a national strategic priority (Obama, 2015).

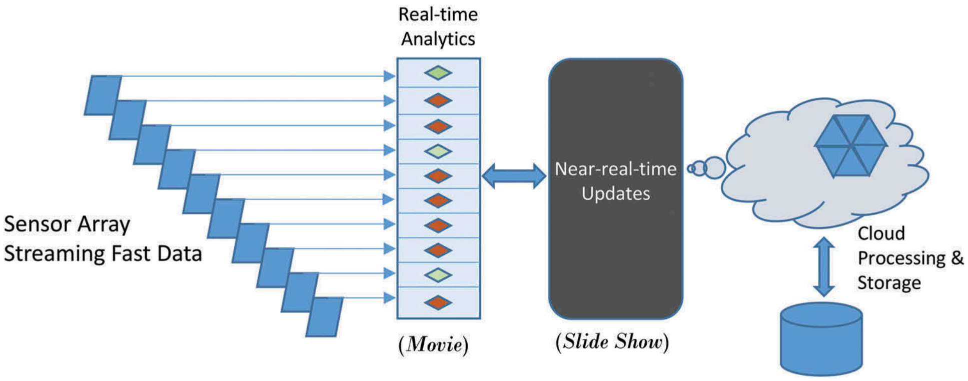


Figure 4. A staged stream processing architecture.

Partly in response to such directives, researchers are intently focused on making advances in the application of high performance computing to key problems. Reed and Dongarra (2015) address specific Strategic Computing objectives and point to advances in the rapidly developing big data analytics computing ecosystem. Big data analytics has emerged out of different business data processing traditions, such as online analytical processing to now include capabilities that employ distributed file systems and parallelism, the most notable being Hadoop (which implements MapReduce). It is expected that computing innovations will continue and that the next milestone of 10^{18} operations (exascale) is approaching rapidly.

Efficiently reaching exascale goals will be difficult, however. Problem areas include power consumption, interprocessor communication latency and big data management as well as the development of a programming model that enables application scalability while being failure and locality aware. As Reed and Dongarra emphasize (2015, p. 65), future systems are expected to have billion-way concurrency where load balancing will become critical. They also point out that arithmetic operations, which are traditional bottlenecks, will become less problematic than memory movement, which will become relatively more expensive. Data locality considerations will reduce communication needs and can be based on an abstract partitioning of geographic space (see Armstrong & Densham, 1992; Wang & Armstrong, 2003).

At the present time, there is a wide gap between powerful supercomputers and the streaming ultralight-weight devices that constitute the IoT. Bridging this gap will require work to connect the leaves to the trunk in a sustainable way. One way to accomplish this task is to integrate parallel infrastructure with the IoT using Cyberinfrastructure (CI). CI is a flexibly configured collection of heterogeneous networked devices (sensors, data repositories, and processing nodes), software, and human resources that are used to address computational and data intensive problems (Wang, 2013, 2016, 2017). The development of CI middleware that coordinates and schedules resources for large geographical problems is challenging because improved performance comes from concurrently executing processes that do not suffer from data starvation and communication latency. Mapping between memory hierarchies and geographic relationships can minimize these difficulties. Wang and Armstrong (2009) suggest that problems be transformed by abstracting them into computebands that are composed of computing "effort" cells that form a tiling that exhausts space, and considers three types of latency and effort: computing time, memory, and I/O. These effort cells are used to guide the allocation of resources to portions of geographically-distributed tasks.

Case study: detecting radiation risk based on cyberGIS and streaming data

We have developed a proof-of-concept to illustrate the challenges of detecting anomalous radioactive sources based on streaming data and to elucidate links between the case study and the four research paradigms discussed earlier. Concepts are illustrated using a cyberGIS workflow designed to detect anomalous radioactivity levels in streaming data anonymized from Safecast (https://blog.safecast.org), a global volunteer-centered citizen science project.

With respect to the first research paradigm (observation and experiment), challenges arise from the perspective of increasing data volume and velocity during the process of radiation data collection. In this case study, the radiation data was collected using a radiation detector (bGeigie Nano) that records incident radiation in counts per second; as the detector approaches a source, the counts increase. Volunteers carry detectors while they are walking or driving, thus forming a dynamic radiation sensor network. More than 3000 volunteers have contributed data, and the size of the Safecast dataset had grown beyond 70 million measurements in June 2017, adding over 2 million measurements monthly (Brown, Franken, Moross, Dolezal, & Bonner, 2017). In addition, 45 Pointcast detectors (fixed real-time detectors) have been installed globally. Each data stream observation includes the following attributes: detector ID, time stamp, location (latitude, longitude), and the measurement strength in counts per second.

The second paradigm represents the theory and models used to analyze radiation streaming data. In this case, two challenges are encountered when detecting anomalous radiation sources using Safecast data. First, nonzero radiation levels may be detected without an anomalous source present, since ionizing radiation occurs naturally (cosmic rays), and is emitted by rocks, soil, and building materials. The challenge is to detect a source with a low signal-to-noise ratio, where the source is the signal, and the background radiation is the ambient noise, in the presence of confounding factors (GPS accuracy, detector motion, shielding, and weather conditions). The second challenge comes from the temporal aspect of the Safecast data. When a radiation source (e.g. nuclear waste) exists in an area as a moving object, the data becomes obsolete quickly if not harnessed to produce a near real-time alarm. Furthermore, Safecast data is voluminous, with high-dimensionality and fast streaming speed, which pose challenges related to the third and fourth research paradigms in the respects of computation- and data-intensive analytics.

To address these challenges, we developed a cyberGIS-enabled analytics framework. Figure 5 illustrates the workflow in the framework where the Safecast data is collected using mobile and fixed sensor networks and transferred to the Safecast central database (paradigm 1). The cyberGIS framework communicates with the Safecast database through Safecast APIs, and the filtered data is stored for analysis in the ROGER supercomputer (paradigm 3 and 4). The CyberGIS-Jupyter environment consists of a Jupyter notebook, Docker containers, cloud-based infrastructure provisioning, and high-performance computing resources, which enables scalable spatiotemporal data query and visualization (Yin et al., 2017). Finally, an algorithm is designed to generate an alarm if an anomalous radioactive source is found (paradigm 2).

The radiation detection algorithm considers two radiation source types: naturally occurring background, and anomalous sources that may come from nuclear weapons, dirty bombs, radioactive waste, or any precursors to such threats. Therefore, we define two types of radiation level estimates (Figure 5). The first estimate is the current radiation level (CRL) that was simulated by randomly selecting one data point in a stream. The second radiation level estimate is the background radiation level (BRL) that is calculated from previous observations in the data stream. To mitigate uncertainties in the radiation data, the BRL at a location should be estimated based on a collection of

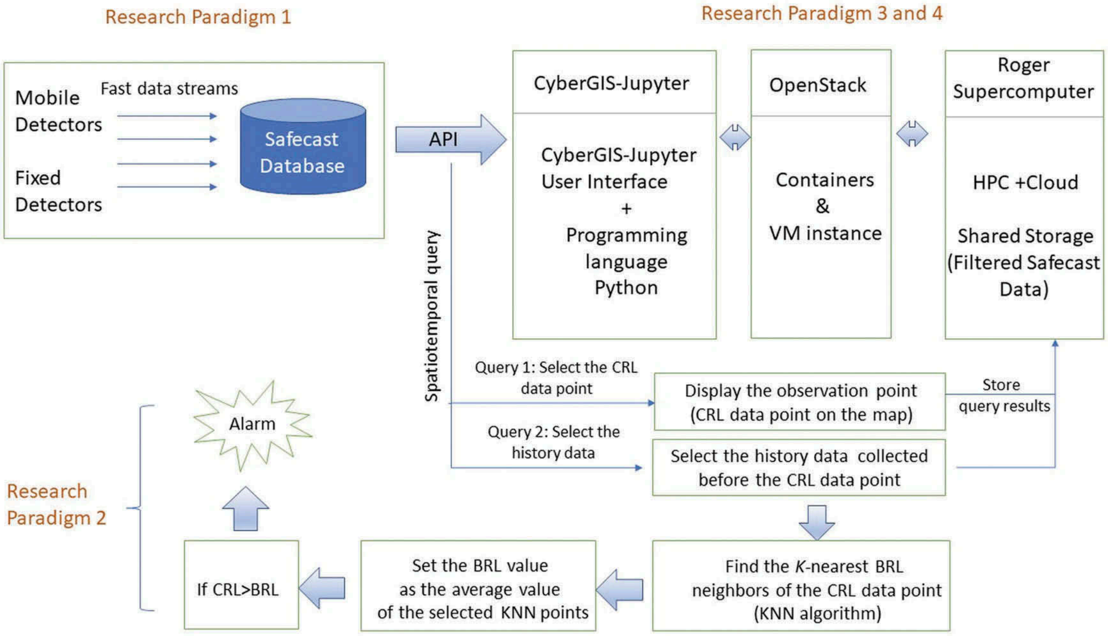

Research Paradigm 1
Research Paradigm 3 and 4
Mobile Detectors
Fast data streams
Fixed Detectors
Safecast Database
API
CyberGIS-Jupyter
CyberGIS-Jupyter User Interface + Programming language Python
OpenStack
Containers & VM instance
Roger Supercomputer
HPC +Cloud
Shared Storage (Filtered Safecast Data)
Spatiotemporal query
Query 1: Select the CRL data point
Display the observation point (CRL data point on the map)
Store query results
Query 2: Select the history data
Select the history data collected before the CRL data point
Alarm
Research Paradigm 2
Find the K-nearest BRL neighbors of the CRL data point (KNN algorithm)
Set the BRL value as the average value of the selected KNN points
If CRL>BRL

Figure 5. CyberGIS workflow for radiation detection.

proximal data points instead of using a single measurement value at a nearby location. Therefore, a k-nearest neighbor (KNN: Keller, Gray, & Givens, 1985) algorithm was implemented and used to estimate the BRL. We first select k-nearest neighbor history data points around the CRL, and those data points are used to calculate the BRL for that particular area as defined by the mean neighborhood radiation value. An alarm will be triggered if the observed CRL exceeds the computed BRL, meaning an anomalous radioactive source is detected.

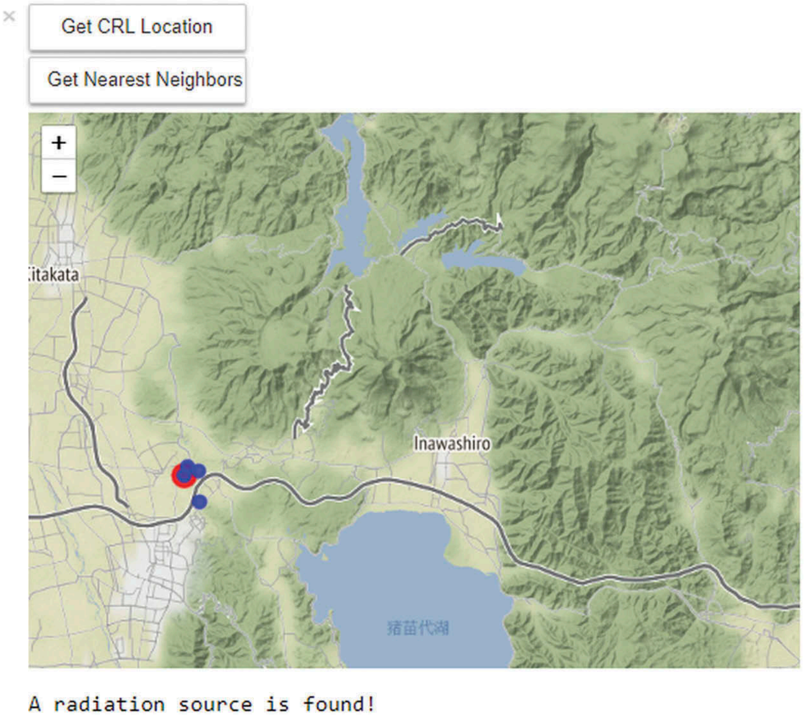
The implementation of the user interface is shown in Figure 6, where the current sensor location is represented as a red circle, and blue circles represent its four nearest neighbors. Since the current sensor radiation level is higher than the mean radiation level of the nearest neighbors, an alarm notification was generated.

Concluding discussion

The proliferation of wirelessly-connected, location-aware devices in the IoT has led to the creation of massive sources of streaming data. These fast data pose a host of challenges to the provision of real-time geographical analysis. These same data also present interesting opportunities to develop new theories and methods of geospatial analysis and knowledge discovery. In particular, as geospatial data collected from mobile sensors are accumulated in a streaming fashion, cyberGIS analytics need to be flexibly adapted to dynamic changes in the volume, pace and spatiotemporal characteristics of data streams. Otherwise, observations will become obsolete, thus diminishing their use in time-sensitive decision-making and knowledge discovery (Wang et al., 2013). The cyberGIS workflow developed in this research is required to resolve the computational intensity associated with KNN search through the use of a high-performance, distributed computing environment. Without cyberGIS and its underlying spatial cyberinfrastructure, this type of scientific workflow would not be possible.

The cyberGIS workflow represents a simple, yet powerful example of geospatial analytics that motivates a rethinking of spatial algorithms and cyberGIS in the context of IoT data streams. Though the KNN algorithm has limitations, advanced alternatives can better resolve data-related fuzziness and uncertainties (Zhang, Demsă, Rantala, & Virrantaus, 2014; Zhang, Demsă, Wang, & Virrantaus, 2018), and also exploit advanced machine learning capabilities (Derrac, García, & Herrera, 2014). CyberGIS analytics for IoT applications will remain challenging for many reasons, with some that can be framed as open challenges and questions that require further research.

- Are their general approaches to achieving real-time performance of computationally intensive spatial algorithms and cyberGIS analytics in data-streaming environments?



A radiation source is found!

Figure 6. User interface for anomalous radiation source detection.

- How are imperfect but "good enough" analytics assessed and reproduced?
- Are there best practices for presenting and communicating imperfect and uncertain outcomes of fast data analytics?

There is no doubt that impactful opportunities for knowledge discovery and geospatial innovation will drive progress on resolving these and other open challenges and questions. Given that most IoT applications have location-based components, it is important to exploit their spatial characteristics in pursuit of scientific advances and benefits to society.

Acknowledgments

The authors wish to thank the reviewers who provided useful comments that led to an improved paper. MPA wishes to thank Ben Speare for his assistance with several figures.

Disclosure Statement

The authors have no financial interest or benefit that has arisen from the direct applications of this research.

Funding

This material is based in part upon work supported by the U.S. National Science Foundation under grant numbers: 1047916, 1443080, and 1743184. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID

Marc P. Armstrong http://orcid.org/0000-0002-5983-7417
Shaowen Wang http://orcid.org/0000-0001-5848-590X
Zhe Zhang http://orcid.org/0000-0001-7108-182X

References

Agarwal, Y., & Dey, A. K. (2016). Toward building a safe, secure, and easy-to-use Internet of things infrastructure. *Computer*, 49(4), 88–91. doi:10.1109/MC.2016.111

Aggarwal, C., Han, J., Wang, J., & Yu, P. (2007). On clustering massive data streams: A summarizing paradigm. In C. Aggarwal (Ed.), *Data streams: Models and algorithms* (pp. 9–38). New York: Springer.

Aggarwal, C., & Yu, P. (2007). A survey of synopsis construction in data streams. In C. Aggarwal (Ed.), *Data streams: Models and algorithms* (pp. 169–207). New York: Springer. doi: 10.1007/978-0-387-47534-9_9

Amdahl, G. M. (1967). Validity of the single-processor approach to achieving large-scale computing capabilities. *Proceedings of the American Federation of Information Processing Societies Conference* (pp. 483–485). Reston, VA: AFIPS Press.

Armstrong, M. P. (1988). Temporality in spatial databases. In *GIS/LIS 88 proceedings: Accessing the world* (Vol. II, pp.880–889). Falls Church, VA: American Society for Photogrammetry and Remote Sensing.

Armstrong, M. P. (2000). Geography and computational science. *Annals of the Association of American Geographers*, 90(1), 146–156. doi:10.1111/0004-5608.00190

Armstrong, M. P., & Densham, P. J. (1992). Domain decomposition for parallel processing of spatial problems. *Computers, Environment and Urban Systems*, Part B. 16 (6), 497–513. doi:10.1016/0198-9715(92)90041-O

Beaumont, J. R. (1989). Towards an integrated information system for retail management. *Environment and Planning A*, 21(3), 299–309. doi:10.1068/a210299

Bennett, D. A., & Armstrong, M. P. (1989). An inductive bit-mapped classifier for terrain feature extraction. In *Proceedings of GIS/LIS '89* (Vol. 1, pp.59–68). Bethesda, MD: American Congress on Surveying and Mapping.

Bennett, D. A., & Armstrong, M. P. (1996). An inductive knowledge based approach to terrain feature extraction. *Cartography and Geographic Information Systems*, 23(1), 3–19. doi:10.1559/152304096782512177

Bennett, D. A., Xiao, N., & Armstrong, M. P. (2004). Exploring the geographic consequences of public policies using evolutionary algorithms. *Annals of the Association of American Geographers*, 94(4), 827–847.doi:10.1111/j.1467-8306.2004.00437.x

Berry, B. J. L. (1964). Approaches to regional analysis: A synthesis. *Annals of the Association of American Geographers*, 54, 2–11. doi:10.1111/j.1467-8306.1964.tb00469.x

Bertino, E., Nepal, S., & Ranjan, R. (2015). Building sensor-based big data cyberinfrastructures. *IEEE Cloud Computing*, 2(5), 64–69.doi:10.1109/MCC.2015.106

Bloom, B. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the Association for Computing Machinery*, 13(7), 422–426. doi:10.1145/362686.362692

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799. doi:10.1080/01621459.1976.10480949

Brown, A., Franken, P., Moross, J., Dolezal, N., & Bonner, S. (2017). The safecast report. https://blog.safecast.org/wp-content/uploads/2017/10/safecastreport2017-part1safecastproject-final-171004011228.pdf

Butler, D. (2014, January 09). Many eyes on Earth. *Nature*, 505, 143–144. https://www.nature.com/polopoly_fs/1.14475!/menu/main/topColumns/topLeftColumn/pdf/505143a.pdf

Calkins, H. W. (1990). Creating large digital files from mapped data. In D. J. Peuquet & D. F. Marble (Eds.), *Introductory readings in geographic information systems* (pp. 209–214). New York: Taylor & Francis.

Calo, R. (2018). Is the law ready for driverless cars? *Communications of the Association for Computing Machinery*, 61(5), 34–36. doi:10.1145/3199599

Cao, H., Mamoulis, N., & Cheung, D. W. (2009). Periodic pattern discovery from trajectories of moving objects. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (2nd ed., pp. 389–408). Boca Raton, FL: CRC Press. https://www.taylorfrancis.com/books/9781420073980/chapters/10.1201%2F9781420073980-19

Chorley, R. J., & Haggett, P. (Eds.). (1967). *Models in geography*. London: Methuen.

Cormode, G., Garofalakis, M., Haas, P., & Jermaine, C. (2012). *Synopses for massive data: Samples, histograms, wavelets and sketches*. Boston: Now Publishers.

Cormode, G., Korn, F., Muthukrishnan, S., & Srivastava, D. (2008). Finding hierarchical heavy hitters in streaming data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(4), Article 16 48. doi:10.1145/1324172.1324174

daCosta, F. (2013). *Rethinking the Internet of Things: A scalable approach to connecting everything*. New York: Springer.

de Selding, P. B. (2015). Facebook-Eutelsat internet deal leaves industry awaiting encore. *SpaceNews*, 26(36), 1, 5.

Deakin, M. (Ed.). (2013). *Smart cities: Governing, modelling and analysing the transition*. New York: Routledge.

Denning, P. J., & Lewis, T. G. (2017). Exponential laws of computing growth. *Communications of the Association for Computing Machinery*, 60(1), 54–65. doi:10.1145/2976758

Derrac, J., García, S., & Herrera, F. (2014). Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects. *Information Sciences*, 260, 98–119.

Dhar, V. (2013). Data science and prediction. *Communications of the Association for Computing Machinery*, 56(12), 64–73. doi:10.1145/2500499

Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M., & Wilson, R. E. (2005). *Mapping crime: Understanding hot spots*. U.S. Department of Justice, National Institute of Justice, NCJ 209393 http://www.ncjrs.gov/pdffiles1/nij/209393.pdf

Eckhoff, D., & Sommer, C. (2014). Driving for big data? Privacy concerns in vehicular networking. *IEEE Security and Privacy*, 12(1), 77–79. doi:10.1109/MSP.2014.2

Ellis, B. (2014). *Real-time analytics: Techniques to analyze and visualize streaming data*. Indianapolis, IN: Wiley.

Fekete, J.-D., & Primet, R. (2016). Progressive analytics: A computation paradigm for exploratory data analysis. https://arxiv.org/abs/1607.05162.

Forrest, S., & Mitchell, M. (2016). Adaptive computation: The multidisciplinary legacy of John H. Holland. *Communications of the Association for Computing Machinery*, 59(8), 58–63. doi:10.1145/2964342

Foust, J. (2015). Orbital debris questions unanswered. *SpaceNews*, 26(23), 1,4.

Gahegan, M. (2000a). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, 32(1), 113–139. doi:10.1111/j.1538-4632.2000.tb00420.x

Gahegan, M. (2000b). The case of inductive and visual techniques in the analysis of spatial data. *Journal of Geographical Systems*, 2(1), 77–83. doi:10.1007/s101090050033

Gahegan, M. (2003). Is inductive machine learning just another wild goose (or might it lay the golden egg)? *International Journal of Geographical Information Science*, 17(1), 69–92. doi:10.1080/713811742

GAO (Government Accountability Office). (2013). In-car location-based services: Companies are taking steps to protect privacy, but some risks may not be clear to consumers. In *GAO-14-81*. Washington, DC: GAO.

Gupta, M., Holloway, C., Heravi, B., & Hailes, S. (2015). A comparison between smartphone sensors and bespoke sensor devices for wheelchair accessibility studies. IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Singapore, 7–9 April. doi:10.1109/ISSNIP.2015.7106900

Hägerstrand, T. (1967). The computer and the geographer. *Transactions of the Institute of British Geographers*, 42, 1–19. https://www.jstor.org/stable/621369

Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, 24(1), 6–21. doi:10.1111/j.1435-5597.1970.tb01464.x

Haggett, P., & Chorley, R. J. (1967). Models, paradigms and the new geography. In *Models in Geography* (pp. 19–41). London: Methuen.

Hegeman, J. W., Sardeshmukh, V. B., Sugumaran, R., & Armstrong, M. P. (2014). Distributed LiDAR data processing in a high-memory cloud-computing environment. *Annals of GIS*, 20(4), 255–264. doi:10.1080/19475683.2014.923046

Helsel, D. R. (2005). *Nondetects and data analysis: Statistics for censored environmental data*. Hoboken, NJ: Wiley.

Henry, C. (2018). Constellation consternation. *SpaceNews*, 29 (4), 13–19.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Holland, J.H., Holyoak, K.J., Nisbett, R.E., & Thagard, P.R. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.

Honda, L., Perkins, J., & Sun, S. (2013). Interpretations of de-orbit, deactivation and shutdown guidelines applicable to GEO satellites. *Proceedings, IEEE Aerospace Conference*, Big Sky, Montana, March 2–9. ISBN: 978-1-4673-1811-2.

IEEE. (1990). *IEEE standard glossary of software engineering terminology* (Standard 610.121990). New York: IEEE

Ilyas, M., Alwakeel, S., Alwakeel, M., & Aggoune, E. (Eds.). (2014). *Sensor networks for sustainable development*. Boca Raton, FL: CRC Press.

Järpe, E. (2002). Surveillance, environmental. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics* (Vol. 4, pp. 2150–2153). New York: Wiley.

Jarr, S. (2015). *Fast data and the new enterprise data architecture*. Sebastopol, CA: O'Reilly

Keller, J., Gray, M., & Givens, J. (1985). A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, 15(4), 580–585. doi:10.1109/TSMC.1985.6313426

Kirkpatrick, K. (2015). The moral challenges of driverless cars. *Communications of the Association for Computing Machinery*, 58(8), 19–20. doi:10.1145/2788477

Kugler, L. (2015). Is "good enough" computing good enough? *Communications of the Association for Computing Machinery*, 58(5), 12–14. doi:10.1145/2742482

Kwan, M. P. (1998). Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis*, 30, 191–217. doi:10.1111/j.1538-4632.1998.tb00396.x

Langran, G., (1988). Temporal design tradeoffs. In *Proceedings of GIS/LIS'88*, Volume 2 (pp. 890–899). Falls Church, VA: ACSM.

Langran, G. (1992). *Time in Geographic Information Systems*. Bristol, PA: Taylor & Francis.

Langran, G., & Chrisman, N. R. (1988). A framework for temporal geographic information. *Cartographica*, 25(3), 1–14. doi:10.3138/K877-7273-2238-5Q6V

Laube, P., & Duckham, M. (2009). Decentralized spatial data mining for geosensor networks. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (2nd ed., pp. 409–430). Boca Raton, FL: CRC Press.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., … Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 119–133. doi:10.1016/j.isprsjprs.2015.10.012

Lochert, C., Scheuermann, B., & Mauve, M. (2010). Information dissemination in VANETs. In H. Hartenstein & K. P. Laberteaux (Eds.), *VANET: Vehicular applications and inter-networking technologies* (pp. 49–80). Chichester, UK: Wiley.

Lowry, I. S. (1968). A short course in model design. In B. J. L. Berry & D. F. Marble (Eds.), *Spatial analysis: A reader in statistical geography* (pp. 53–64). Englewood Cliffs, NJ: Prentice-Hall.

Mandl, K., Overhage, J., Wagner, M., Lober, W., Sebastiani, P., Mostashari, F., … Grannis, S. (2004). Implementing syndromic surveillance: A practical guide informed by early experience. *Journal of the American Medical Informatics Association*, 11(2), 141–150. doi:10.1197/jamia.M1356

Marble, D. F. (2015). Computational geography as a new modality. In M. Monmonier (Ed.), *The history of cartography* Volume 6, Part 1. Chicago: University of Chicago Press. 488–492.

Marble, D. F., & Anderson, B. M. (1972). *LANDUSE: A computer program for laboratory use in economic geography courses*. Commission on College Geography Technical Paper No. 8. Washington, DC: Association of American Geographers.

Miller, H. J. (1991). Modeling accessibility using space-time prism concepts within a GIS. *International Journal of Geographical Information Systems*, 5, 287–301. doi:10.1080/02693799108927856

Miller, H. J. (2005). A measurement theory for time geography. *Geographical Analysis*, 37, 17–45. doi:10.1111/j.1538-4632.2005.00575.x

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, 50, 181–201. doi:10.1111/j.1467-9787.2009.00641.x

Miller, H. J., & Han, J. (2009). Geographic data mining and knowledge discovery: An overview. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (2nd ed., pp. 1–26). Boca Raton, FL: CRC Press.

Mohring, K., Myers, T., Atkinson, I., VanDerWal, J., & van der Valk, S. (2015). Sensors in heat: A pilot study for high resolution urban sensing in an integrated streetlight platform. International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Singapore, 7–9 April. doi:10.1109/ISSNIP.2015.7106908

Mone, G. (2015). The new smart cities. *Communications of the Association for Computing Machinery*, 58(7), 20–21. doi:10.1145/2771297

Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8), 114–117. doi:10.1109/N-SSC.2006.4785860

Moreau, T., Sampson, A., & Ceze, L. (2015). Approximate computing: Making mobile systems more efficient. *IEEE Pervasive Computing*, 14(2), 9–13. doi:10.1109/MPRV.2015.25

Narkhede, N., Shapira, G., & Palino, T. (2017). *Kafka: The definitive guide*. Sebastopol, CA: O'Reilly.

NAS (National Academies of Sciences, Engineering, and Medicine). (2016). *Future directions for NSF advanced computing infrastructure to support U.S. Science and engineering in 2017–2020*. Washington, DC: The National Academies Press. doi:10.17226/21886

Nayak, A., & Stojmenovic, I. (2010). *Wireless sensor and actuator networks: Algorithms and protocols for scalable coordination and data communication*. Hoboken, NJ: Wiley.

NIST (National Institute of Standards and Technology). (2015). *DRAFT NIST big data interoperability framework: Volume 1, definitions*. NIST Special Publication 1500-1. Gaithersburg, MD: NIST. doi:10.6028/NIST.SP.1500-1

Noor, T. H., Sheng, Q. Z., Maamar, Z., & Zeadally, S. (2016). Managing trust in the cloud: State of the art and research challenges. *Computer*, 49(2), 34–45. doi:10.1109/MC.2016.57

NRC (National Research Council). (2013). *Frontiers in massive data analysis*. Washington, DC: The National Academies Press. doi:10.17226/18374

Obama, B. (2015). https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

Openshaw, S. (1995). Developing automated and smart spatial pattern exploration tools for geographical information systems applications. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 44(1), 3–16. doi:10.2307/2348611

Openshaw, S. (1998). Towards a more computationally minded scientific human geography. *Environment and Planning A*, 30, 317–332. doi:10.1068/a300317

Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3), 441–462. doi:10.1111/j.1467-8306.1994.tb01869.x

Peuquet, D. J., & Duan, N. (1995). An event-based spatio-temporal data model (ESTDM) for temporal analysis of geographical data. *International Journal of Geographical Information Systems*, 9(7), 7–24. doi:10.1080/02693799508902022

Porter, M. E., & Heppelmann, J. E. (2014). How smart connected devices are transforming competition. *Harvard Business Review*, (Nov.), 70–86. https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

Rajasegarar, S., Zhang, P., Zhou, Y., Karunasekera, S., Leckie, C., & Palaniswami, M. (2014). High resolution spatio-temporal monitoring of air pollutants using wireless sensor networks. Proceedings, IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Singapore, 21–24 April. doi:10.1109/ISSNIP.2014.6827607

Reed, D. A., & Dongarra, J. (2015). Exascale computing and big data. *Communications of the Association for Computing Machinery*, 58(7), 56–68. doi:10.1145/2699414

Rogerson, P. A. (2009). Monitoring changes in spatial patterns. In S. A. Fotheringham & P. A. Rogerson (Eds.), *The SAGE handbook of spatial analysis* (pp. 343–354). Thousand Oaks, CA: Sage.

Romero, J. (2015). On New York's Lake George, researchers fire up a state-of-the-art observatory. http://www.sciencemag.org/news/2015/07/new-york-s-lake-george-researchers-fire-state-art-observatory doi:10.1126/science.aac8828

Rubinfeld, R., & Shapira, A. (2011). Sublinear time algorithms. *SIAM Journal of Discrete Mathematics*, 25(4), 1562–1588. doi:10.1137/100791075

Serpanos, D. (2018). The cyber-physical systems revolution. *IEEE Computing Edge*, 4(5), 10–13. doi:10.1109/MC.2018.1731058

Shaw, S.-L. (2006). What about 'time' in transportation geography? *Journal of Transport Geography*, 14(3), 237–240. doi:10.1016/j.jtrangeo.2006.02.009

Shekhar, S., Evans, M. R., Gunturi, V., Yang, K., & Cugler, D. C. (2014). Benchmarking spatial big data. In T. Rabl, M. Poess, C. Baru, & H. A. Jacobsen (Eds.), *Lecture notes in computer science: Vol. 8163. Specifying big data benchmarks* (pp. 81–93). Berlin: Springer.

Shi, W., & Dustdar, S. (2016). The promise of edge computing. *Computer*, 49(5), 78–81. doi:10.1109/MC.2016.145

Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., de Carvalho, A. C., & Gama, J. (2013). Data stream clustering: A survey. *ACM Computing Surveys*, 46(1), Article 13. doi:10.1145/2522968.2522981

Struchen, R., Vial, F., & Andersson, M. G. (2017). Value of evidence from syndromic surveillance with cumulative evidence from multiple data streams with delayed reporting. https://www.nature.com/articles/s41598-017-01259-5 doi:10.1038/s41598-017-01259-5

Thrift, N. (1977). An introduction to time geography. In *Concepts and techniques in modern geography (CATMOG) No 13*. Norwich, UK: Geo Abstracts.

Tomlinson, R. F. (Ed.). (1970). *Environment information systems*. Williamsville, NY: International Geographical Union Commission on Geographical Data Sensing and Processing.

Torrens, P. M. (2010). Geography and computational social science. *GeoJournal*, 75(2), 133–148. doi:10.1007/s10708-010-9361-y

Tsai, V. J. D. (1993). Delaunay triangulations in TIN creation: An overview and a linear-time algorithm. *International Journal of Geographical Information Science*, 7(6), 501–524. doi:10.1080/02693799308901979

Turkay, C., Kaya, E., Balcisoy, S., & Hauser, H. (2017). Designing progressive and interactive analytics processes for high-dimensional data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1), 131–140. doi:10.1109/TVCG.2016.2598470

Vilajosana, I., Llosa, J., Martinez, B., Domingo-Prieto, M., Angles, A., & Vilajosana, X. (2013, June). Bootstrapping smart cities through a self-sustainable model based on big data flows. *IEEE Communications Magazine*, pp. 128–134. doi:10.1109/MCOM.2013.6525605

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1), 37–57. doi:10.1145/3147.3165

Wachowicz, M. (1999). *Object-oriented design for temporal GIS*. London Taylor & Francis.

Wang, S. (2013). CyberGIS: Blueprint for integrated and scalable geospatial software ecosystems. *International Journal of Geographical Information Science*, 27(11), 2119–2121. doi:10.1080/13658816.2013.841318

Wang, S. (2016). CyberGIS and spatial data science. *GeoJournal*, 81(6), 965–968. doi:10.1007/s10708-016-9740-0.pdf

Wang, S. (2017). CyberGIS. In D. Richardson, N. Castree, M. F. Goodchild, A. L. Kobayashi, W. Liu, & R. Marston (Eds.), *The international encyclopedia of geography: People, the earth, environment, and technology*. Hoboken, NJ: Wiley. doi:10.1002/9781118786352.wbieg0931.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M. F., Liu, Y., & Nyerges, T. L. (2013). CyberGIS software: A synthetic review and integration roadmap. *International Journal of Geographical Information Science*, 27(11), 2122–2145. doi:10.1080/13658816.2013.776049

Wang, S., & Armstrong, M. P. (2003). A quadtree approach to domain decomposition for spatial interpolation in grid computing environments. *Parallel Computing*, 29(10), 1481–1504. doi:10.1016/j.parco.2003.04.003

Wang, S., & Armstrong, M. P. (2009). A theoretical approach to the use of cyberinfrastructure in geographical analysis. *International Journal of Geographical Information Science*, 23(2), 169–193. doi:10.1080/13658810801918509

Wang, W., & Wets, G. (Eds.). (2013). *Computational Intelligence for Traffic and Mobility*. Paris: Atlantis.

Wang, X. (2018). Data acquisition in vehicular ad hoc networks. *Communications of the Association for Computing Machinery*, 61(5), 83–88. doi:10.1145/3197544

Weinberg, S. (2015). *To explain the world: The discovery of modern science*. New York: HarperCollins.

Wiener, P., Stein, M., Seebacher, D., Bruns, J., Frank, M., Simko, V., … Mimis, J. (2016). BigGIS: A continuous refinement approach to master heterogeneity and uncertainty in spatio-temporal big data (vision paper). *GIS '16 Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Article No. 8. doi:10.1145/2996913.2996931

Worboys, M. F. (1994). A unified model for spatial and temporal information. *The Computer Journal*, 37(1), 26–34. doi:10.1093/comjnl/37.1.26

Wu, H., Guan, X., & Gong, J. (2011). ParaStream: A parallel streaming Delaunay triangulation algorithm for LiDAR points on multicore architectures. *Computers and Geosciences*, 37(9), 1355–1363. doi:10.1016/j.cageo.2011.01.008

Xiong, D., & Marble, D. F. (1996). Strategies for real-time spatial analysis using massively parallel SIMD computers: An application to urban traffic flow analysis. *International Journal of Geographical Information Systems*, 10(6), 769–789. doi:10.1080/02693799608902109

Yan, J., Cowles, M. K., Wang, S., & Armstrong, M. P. (2007). Parallelizing MCMC for Bayesian spatiotemporal geostatistical models. *Statistics and Computing*, 17(4), 323–335. doi:10.1007/s11222-007-9022-2

Yang, C., Xu, Y., & Nebert, D. (2013). Redefining the possibility of digital Earth and geosciences with spatial cloud computing. *International Journal of Digital Earth*, 6(4), 297–312. doi:10.1080/17538947.2013.769783

Yang, C., Yu, M., Jiang, Y., & Li, Y. (2017). Utilizing cloud computing to address big geospatial data challenges. *Computers, environment and urban systems*, *61*, 120–128. doi:10.1016/j.compenvurbsys.2016.10.010

Yin, D., Liu, Y., Padmanabhan, A., Terstriep, J., Rush, J., & Wang, S. (2017). A cyberGIS-jupyter framework for geospatial analytics at scale. In *Proceedings of the Practice and Experience in Advanced Research* Computing *2017 on Sustainability, Success and Impact*, 18:1–18: 8. New York: ACM. doi:10.1145/3093338.3093378

Yuan, M. (2009). Toward knowledge discovery about geographic dynamics in spatiotemporal databases. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (2nd ed., pp. 347–365). Boca Raton, FL: CRC Press.

Yuan, M., & Stewart, K. (Eds.). (2007). *Computation and visualization for understanding dynamics in geographic domains: A research agenda*. Boca Raton, FL: CRC Press.

Zhang, H., Chen, G., Ooi, B., Tan, K., & Zhang, M. (2015). In-memory big data management and processing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 27(7), 1920–1948. doi:10.1109/TKDE.2015.2427795

Zhang, Z., Demsǎ, U., Rantala, J., & Virrantaus, K. (2014). A fuzzy multiple-attribute decision making modelling for vulnerability analysis on the basis of population information for disaster management. *International Journal of Geographical Information Science*, 28(9), 1922–1939. doi:10.1080/13658816.2014.908472

Zhang, Z., Demsǎ, U., Wang, S., & Virrantaus, K. (2018). A spatial fuzzy influence diagram for modelling spatial objects' dependencies: A case study on tree-related electric outages. *International Journal of Geographical Information Science*, 32(2), 349–366. doi:10.1080/13658816.2017.1385789