

ePub^{WU} Institutional Repository

Laurie A. Schintler and Manfred M. Fischer

The Analysis of Big Data on Cites and Regions - Some Computational and Statistical Challenges

Paper

Original Citation:

Schintler, Laurie A. and Fischer, Manfred M. (2018) The Analysis of Big Data on Cites and Regions - Some Computational and Statistical Challenges. *Working Papers in Regional Science*, 2018/08. WU Vienna University of Economics and Business, Vienna.

This version is available at: <http://epub.wu.ac.at/6637/>

Available in ePub^{WU}: November 2018

ePub^{WU}, the institutional repository of the WU Vienna University of Economics and Business, is provided by the University Library and the IT-Services. The aim is to enable open access to the scholarly output of the WU.

The Analysis of Big Data on Cities and Regions – Some Computational and Statistical Challenges

Laurie A. Schintler

Associate Professor

Schar School of Policy and Government,

George Mason University, USA

lschintl@gmu.edu

Manfred M. Fischer

Professor Emeritus of Economic Geography

Department of SocioEconomics

Vienna University of Economics and Business

Vienna, Austria

manfred.fischer@wu.ac.at

Abstract. Big Data on cities and regions bring new opportunities and challenges to data analysts and city planners. On the one side, they hold great promise to combine increasingly detailed data for each citizen with critical infrastructures to plan, govern and manage cities and regions, improve their sustainability, optimize processes and maximize the provision of public and private services. On the other side, the massive sample size and high-dimensionality of Big Data and their geo-temporal character introduce unique computational and statistical challenges. This chapter provides overviews on the salient characteristics of Big Data and how these features impact on paradigm change of data management and analysis, and also on the computing environment.

Key Words: massive sample size, high-dimensional data, heterogeneity and incompleteness, data storage, scalability, parallel data processing, visualization, statistical methods

JEL Classifiers: C10, C55, C80

1 Introduction

Over the past two decades, we have seen a paradigm shift in the way information and data is generated and handled. This shift is driven by several factors: (i) the significant improvements in storage capacity and computing power to process very large data sets; (ii) the rapid increase in remote sensors generating new streams of digital data from telescopes, traffic monitors and video cameras monitoring the environment; (iii) the introduction of the Internet of Things, implying that even simple components, sensors, and devices can communicate over the internet; (iv) the mobile revolution with the advent of location-enabled communications devices such as smartphones, enabling to receive and send information anytime and everywhere; (v) the emergence of electronic commerce channels and social media platforms; and (vi) crowdsourcing platforms for volunteered geographic information (VGI), a type of user-generated content with a geospatial component. These changes together have resulted in what is called Big Data.

Big Data on cities and regions bring new opportunities and challenges to data analysts and city planners. On the one hand, they hold great promise to combine increasingly detailed and personalized data for each citizen with critical infrastructures to plan, govern and manage cities and regions, improve their sustainability, optimize processes and maximize the provision of public and private services. On the other hand, Big Data introduce unique computational and analytical challenges, which compromise its value in smart city or regional contexts.

A number of challenges in both data management and data analysis call for new strategies and solutions to support the Big Data era. Major challenges that arise include: (i) handling different data formats and structures, (ii) dealing with the massive and high-dimensional nature of the data, (iii) developing algorithms that exploit parallel and distributed architectures, (iv) coping with sample biases and heterogeneity, (v) developing methods for visualizing massive data, and (vi) coping with the need for real-time analysis and decision making. These challenges are magnified in cases where Big Data are distributed across locations (National Research Council, 2013). In addition, spatio-temporal Big Data create specific problems and difficulties that need to be taken into consideration.

This chapter provides overviews on the salient characteristics of Big Data and how these features impact on changing the classical paradigm of data management and analysis, and also the computing environment. In doing so the contribution presents a broad and encompassing overview of the topic, highlighting what we consider to be the main challenges and tasks for the future. Major issues, such as those relating to data ownership, education and training, costs, security, epistemology, policy, law and ethics – while important topics in their own right – are beyond the current scope. Thus, we set such topics aside to keep the discussion appropriately focused.

2 Big Data and Prospects for Research

2.1 What is Big Data?

The term Big Data has been widely used for any sort of data sets that is larger than usual. Big Data has features that are not shared by traditional data sets. In essence, there are three aspects that define Big Data (Miller, 2016; Pattnaik and Mishra, 2016).

Volume: Big Data is about size – massive volumes of data beyond the capability of traditional approaches of data analytics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information sensory mobile devices such as aerial sensory technologies, radio-frequency readers, cameras, and wireless sensor networks. Unlike in traditional data sets, Big Data are characterized not only by massive sample size, but also high-dimensionality. Much of the data is geographic in nature containing explicit or implicit spatial information. Terabyte archives for remotely sensed imagery data, vast volumes of real-time sensor observations and location-based media data, and VGI data are examples where new innovative procedures for handling and analyzing massive volumes of spatial data have been or still have to be developed.

Velocity: Big Data is generated in a very rapid pace, and often has to be processed quickly. Traffic data in mobile communication networks and streaming video data are prime examples. The velocity of Big Data is also relevant in the Internet of Things where an up-to-date picture of information and near real-time responses are prerequisites.

Variety: Big Data is highly heterogeneous in nature. Data come from multiple data sources, and the level in which they are structured tends to vary from data source to data source. Data increasingly lose structure and many new formats that occur go beyond relational data bases, pure text, photo, video, web and GPS (global positioning system) data. Transferring unstructured data into structured format for later analysis is a major challenge.

There are more aspects that recently came into view to support the above three V's (volume, velocity and variety) of Big Data to further define Big Data, and these are 'veracity' and 'value'. Veracity relates to uncertainty of data and data incompleteness, while value to the aspect to turn Big Data into values otherwise useless (Pattnaik and Mishra, 2016).

Big Data are complex in a variety of ways. They are voluminous, high-dimensional, heterogeneous, multi-source and collected over a range of temporal and spatial scales. Spatial data may come from earth observations, social media, mobile phone calls, and unmanned aerial vehicles. Sensor technology is also being embedded in vehicles and containers, adding to the abundance of data. Moreover, the deployment of the Internet of Things will produce large amounts of text-like communication between devices, people, and places.

Through the whole spectrum of society and business, vast volumes of data are collected on our physical and human-made environment, including building structures, nightlights, land use cover, meteorological conditions, water quality, and so on. Large-scale simulations based on this data (e.g., global climate modeling) provide an additional layer of data in Geographical Information Systems (GISs). The world wide web, and complex ecosystems of online electronic commerce websites and infomediaries (e.g., job markets, dating websites, recommendation services), repositories of digitized documents, open data portals, social media platforms, and other websites it encompasses, give us a rich and unfolding picture of the interests, preferences, needs, and activities of individuals, organizations, and firms in cities and regions all over the

world. Web 2.0 or the interactive web and related social media platforms, ‘apps’, and discussion fora, in particular, have created a new generation of sensors, namely humans (or citizens) as sensors. Mobile devices including smartphones and location acquisition technologies such as global position systems are producing realms of spatial trajectory data that capture detailed information on human, material, and information, and animal movements.

Emerging technologies, such as blockchain, nanotechnology, cloud robotics, are contributing to even newer sources of Big Data. Such cutting-edge technologies are also advancing our capacity to store, process, and glean insight and knowledge from Big Data. The Internet of Things, which comprises a large and growing assemblage of interconnected devices, is actively monitoring and intelligently processing everything from the contents of our refrigerators, for example, to the second-to-second operational characteristics of large-scale infrastructure. Cyber-physical systems, which integrate computing, networking, and physical technologies in a complex and adaptive fashion, are a burgeoning source of Big Data. For example, automated vehicles collect vast amounts of real-time data about traffic conditions and other aspects of the surrounding environment, information that is instantaneously fed back to the cloud for processing to optimize vehicular routing and performance. Indeed, machine-generated data – that is, raw data produced by machines – is a rapidly expanding type of data. Such machine-generated data could soon make up 50 percent of all of the data in the world (Gantz and Reinsel, 2012).

Just like a-spatial Big Data, geospatial Big Data (or Big Spatial Data) contains disparate formats, structures, semantics, granularity, and so forth. However, space and time dimensions of the data add further heterogeneity. To this point, spatial data comprises varying spatial and temporal scales, levels of resolution, and extents of coverage, and with different spatial referencing systems (Fischer, Scholten and Unwin, 1996). Citizen sensing, crowd-sourcing and other forms of user-generated data tend to have a high degree of spatial and temporal resolution – that is, information that is often summarized down to latitude and longitude coordinates, and seconds of the day – and coverage that extends over the entire globe. Other types of spatial data, such as those collected from official organizations are more aggregated, and limited in geographic scope. The heterogeneity of Big Data also stems from the particular characteristics of the data acquisition devices themselves. Sensors are either positioned on moving objects or static, continually monitoring the changing environment in an area or at a particular location (Li et al., 2016). Thus, spatial objects are classified geometrically as line, point, or area (Fischer and Wang, 2011).

Big Spatial Data is fraught with heterogeneity, but also with noise, incompleteness, redundancy, uncertainty, and other undesirable features. For example, sensors that monitor the environment produce repetitive coverage, since multiple images must be collected in a short amount of time to achieve appropriate and adequate spatial coverage. Mobile trace data tends towards noise and incompleteness, given that location positioning technologies are currently unable to produce proper signals in specific environments. Crowd-sourced geographic information data often contain duplicate records stemming from human error, technological and algorithmic glitches (Kwan, 2016). Moreover, user-generated data is notoriously biased towards demographic characteristics, preferences, interests, and activity patterns of their users. The digital divide is a

further source of bias and gaps in Big Data (Schintler, 2017). Given that regions have different demographic, economic, cultural, and technological profiles, the type and extent of bias vary from place to place.

2.2 The Promise of Big Data

To the extent that the challenges surrounding Big Data analysis can be effectively managed, the hope is that the frontiers of science will expand in transformative ways, and technology will become more adaptive, flexible, personalized, and robust (National Research Council, 2013). Big Data offers enormous opportunities for cities and regions, especially in the era of the ‘new urban world’ (Kourtit and Nijkamp, 2018). Specifically, new and expanding sources of data, coupled with advanced computational and analytical methods and techniques, can enable communities to operate as high-powered cognitive engines (Batty, 2013).

Moreover, new models of computation, methods and techniques that combine data, simulations, predictive analytics, and visualization can help in better understanding cities and regions in the first place. In the era of Big Data then, it is easy to imagine cities and regions in which increasingly refined and customized data are collected and maintained for each citizen and in which such data are combined with critical infrastructures (including not only buildings and streets, but also water, gas and electricity pipelines) to plan, govern, manage, and control cities and regions in an optimal manner, to ultimately enhance their sustainability, livability, and competitiveness.

Understandably, there is a great deal of optimism about the potential of Big Data. Recent advancements in computer hardware (faster CPUs, cheaper memory), and new technologies and software for processing Big Data have made it easier to collect, analyze and mine massive amounts of structured and unstructured data. Indeed, our knowledge of how to design scalable data-centric technologies through cloud computing and storage, and parallel and distributed platforms, tailored to the nuances of Big Data, has been greatly enhanced in the last few decades. In addition, innovations in the fields of machine learning, statistics, and algorithmic theory have produced analytical methods that can handle increasingly large and multi-source data sets (National Research Council, 2013).

3 Challenges

The prospects of Big Data, however, must be appropriately balanced by an understanding of the major difficulties and challenges that conflict with the envisioned aims of Big Data in science and society. The massive sample size and high-dimensionality of Big Data pose unique and profound computational and statistical challenges. This section briefly discusses challenges that arise in the five distinctive stages of the data analysis pipeline that leads from ‘data acquisition and recording’ over ‘information extraction and cleaning’ and ‘data integration, aggregation and representation’ to ‘query processing, modeling and analysis’ and ‘visualization and interpretation’.

3.1 Data acquisition and recording

Managing large volumes and varieties of (structured and unstructured) data is one of the most apparent technical challenges in Big Data analytics. Big Data is first acquired from some generating source (or sources) and then transmitted to storage and recorded for future use. In some cases, it may not be a viable option to store data. This can be either because it is physically and/or economically not feasible to store immensely large volumes of data, or because management overhead becomes too large, for example, when data is updated faster than it can be stored. Streaming processing is an approach to address several of the challenges concerning the analysis of the data that cannot be, may not be, or is better not to store (Andrada, Gedik and Turaga, 2014).

Because of the large size of high-dimensional data, it is often necessary to use compressed data instead. Lossy compression is the class of data encoding methods often used to reduce data size for storage and handling. While lossy compression is effective in reducing the volume of Big Data, it comes at the cost of information loss caused by inexact approximations and partial data discarding. Information loss is especially problematic in the case of data produced by multiple sensors of different types. For Big Spatial Data, in particular, information on spatial relations and generalization can be lost in the compression. Instead one can conduct dimensionality reduction – another technique used in the high-dimensional data acquisition and processing context – that represents high-dimensional data points in a lower-dimensional space while preserving hereby properties of the data as much as possible (see van der Maaten, Postma and van den Herik, 2009). However, such processes are computationally intensive, particularly in the case of space-time data. Use of clustering algorithms explicitly designed for spatial (and spatiotemporal) data – e.g., the Spatio-Temporal Density-based Spatial Clustering of Applications with Noise (ST-DBSCAN) algorithm – may help in managing this problem (Li et al., 2016).

An increasing number of social media and mobile technologies are generating geo- and time-tagged data. In such cases, it is common to develop streaming algorithms that attempt to process the data in real-time, avoiding storage. Examples include early alert systems for disease outbreaks. The requirement for real-time processing creates new algorithmic challenges, where ‘answer quality’ needs to be traded off against ‘answer timeliness’. Many data sets are also indexed by spatial coordinates. This yields new algorithmic challenges, where ‘answer quality and timeliness’ has to be traded off against the geographic granularity of the answer (National Research Council, 2013).

When dealing with data sets from diverse sources, systematic recording and tracking of data quality metadata are very important. Metadata is used to record information about the data, for example, sample size, sampling strategy, scale, availability, age, ownership, and price (if relevant) (Getis, 1999). However, creating metadata for Big Data is complicated and often impractical. One challenge is that Big Data tends to change hands frequently, where it gets repurposed, repackaged, and reprocessed at each stop (Schintler and Chen, 2017). Thus, details of the data often get lost as it travels from one person or organization to another. Moreover,

attributes are sometimes hidden, as is often the case with proprietary or personally-sensitive Big Data (Getis, 1999). In crowd-sourced or user-generated data, information on the granularity of data in space and time and related details are often missing, making full and proper documentation of such data difficult (Li et al., 2016). The ability for automatically generating the metadata, however, is currently underdeveloped.

3.2 Information extraction and cleaning

Frequently raw data collected will not be in a format ready for analysis. For example, we must convert unstructured data in the form of text to structured data before it is suitable for using classical modeling and analysis tools. In the case of geospatial data, it requires geocoding before using it in a GIS. We expect an information extraction process that pulls out the required information from the underlying sources and expresses it in a standard form appropriate for the analysis at hand. Doing this correctly is a continuing technical issue. In a city or regional environment, we need to be able to extract information on the location of features, and the spatial context of these objects from the data. While some sources of Big Data contain explicit geographic references – e.g., latitude and longitude coordinates – many others do not. For example, in social media data geographic information is embedded in the feeds, often across multiple rather than a single entity, and the information is in poorly-defined formats (Vatsavai et al., 2012).

Most sources of data are far from perfect. Data tend to be corrupted by either systematic bias or random noise, or both. Measurement processes are a major source of noise, as are data generated from simulations which hinge on the underlying quality of the initial data in the first place (National Research Council, 2013). Noise and spurious correlation are especially problematic in the case of Big Data given the high-dimensionality of the data (Fan, Han and Liu, 2014). Having a larger sample size does not necessarily mitigate against these problems. Even data obtained by high-quality instrumentations or through robust sampling can be problematic (National Research Council, 2013).

The practice of cleaning data is fairly well established for small and moderate data sets, but new challenges arise in the context of large samples (Osborne, 2013). While the tasks of detecting mistakes, missing information, and other imperfections in samples of small data – for example, through sanity checking and data exploration techniques – can be applied in the case of Big Data, finding representative samples in large data sets poses challenges in this regard. Moreover, human interaction is impossible in such situations, given time limitations and the size of the data. Thus, it is desirable to have automatic cleaning mechanisms embedded into the data acquisition and data storage software (National Research Council, 2013).

3.3 Data integration, aggregation, and representation

In general, the value of data increases, when linked with other data. Hence, data integration can act as a useful means to create value. But integration of Big Data collected from different sources

is difficult due to the diversity of data types and formats, semantics, ownership, organizational structures and levels of resolution, and so on. It is even more complicated in the case of Big Spatial Data, given the varying spatial/temporal scales, levels of granularity and coverage the data comprises (Fischer, Scholten and Unwin, 1996). Data aggregation in spatial and temporal data is fraught by the modifiable areal unit and the modifiable temporal unit problems (Manley, 2014). Spatial data magnifies these issues, given that there are countless ways to parse and aggregate the data spatially and temporally.

Big Data on cities and regions are typically highly distributed and generally remain distributed because of technical, political, social and economic reasons. Due to limitations in transmitting massive volumes through channels with limited bandwidth, the highly distributed nature of Big Data creates challenges in terms of data access, integration, and sharing. Moreover, not all the data produced by different sources are defined using the same data representation techniques, and this imposes additional challenges in terms of managing data in a distributed environment. Data representation involves selecting an appropriate mathematical structure with which to model the data to reduce computation in a way that leads not only to algorithmically efficient, but also statistically meaningful results (National Research Council, 2013).

3.4 Query Processing, Modeling, and Analysis

One basic operation in processing Big Data involves querying part of the data, which is usually done by indexing. Different types of queries and data require different indexing methods to ensure that search and retrieval of information is efficient and effective (National Research Council, 2013). For spatial queries – such as ‘find all features located in a particular region’ or ‘find all objects that contain a given query point’ – processing is computationally intensive because of the polynomial complexity of the geometric operations required to pull data. In multidimensional data, there are additional spatial relationships, which further impede the efficiency of query processing (Wang, Aji and Vo, 2015).

Recent research on spatial query processing of real-time streaming Big Data focuses on designing indexing methods, which segment the search space into tiles, such that search time focuses on a single tile at a time. However, an open question is how to organize the tiles in such a way that the search process is efficient. Hilbert space-filling curves may help in addressing this concern (Li et al., 2016). When querying Big Spatial Data, we also need to ensure adequate extracting and appropriate samples from the data, as failure to do so increases the probability of erroneous conclusions. This is a problem with immense spatial data as there are many possible realizations that can be drawn from a single source (Getis, 1999).

The primary goal of analyzing Big Data is to derive knowledge from data. For achieving this, typically statistical models are used as a convenient framework. Statistical models allow to identify relationships between variables and to understand how these variables impact on the system of interest. Statistical models, moreover, enable one to make predictions along with coverage intervals reflecting uncertainties. Although parametric statistical models may play an important role in data analysis, especially in contexts where the model can be – at least partly –

specified from an underlying theory, the nonparametric perspective is more in line with the exploratory and predictive goals of Big Data analysis (National Research Council, 2013).

A large body of methods exists for small-scale to medium-scale data analysis and machine learning (notably data mining), but most are difficult or impossible to use for massive and high-dimensional data because they do not interface well with existing large-scale computer systems and architectures, such as multi-core processors or distributed clusters of machines. Hence, a major challenge in large-scale data representation is to extend work developed in the context of single machines and medium-scale data to parallel, distributed processing and much larger-scale contexts. In a distributed computing environment Bayesian models with parameters estimated using Markov Chain Monte Carlo simulation can relatively easily make advantage of multiple computers by performing independent simulations (National Research Council, 2013).

Machine learning – based on well-grounded statistical principles and coupled with reinforcement learning, support vector machines, Bayesian networks, evolutionary and swarm-based algorithms – provides important means to read out value from data (see Bishop, 2006; Panda, Dehuri and Patra, 2015; Raschka and Mirjalili, 2017). The problem of massive sample size and high-dimensionality of data is generally solved through parallelization of algorithms accomplished either by data parallelism or task parallelism. Machine learning methods can capture nonlinearity, heterogeneity, noise, and other complexities in spatial and temporal data. Feedforward neural networks, in particular, may be used for nonparametric statistical inference, as they do not require a priori specifications of functional forms to be approximated (Fischer, 2015). Deep learning, an emerging paradigm within machine learning (see Bengio, 2009), focuses on features composed of multiple levels of nonlinear operations, such as in feedforward neural networks with many hidden layers. But searching the parameter space of deep architectures is a challenging task.

3.5 Visualization and Interpretation

Visualization and interaction technologies may give users a gateway into their – massive amounts of structured and unstructured – data. Visualization can, for example, be used to uncover hidden patterns and spot outliers, which can reveal ways in which the data could be better partitioned for further computational analysis. Systems with a rich palette of visualization tools become essential in conveying to the users the results of the queries in a way that is best understood in the particular domain (Miller, 2016). Ultimately, display of Big Data appears to be useful but only if succinctly and correctly summarizing the underlying information. Related to this, the user should be able to easily and quickly scrutinize each piece of data that she sees, to learn to know its provenance, which is critical to understanding the data in the first place.

For smart cities and regions, the challenge is to design visualization tools that enable policy and decision makers, city and regional planners, and the community-at-large to visually explore and analyze the data for better decision making (Li et al., 2016). Dashboards and geoportals have great utility in this context (Batty et al., 2012). More research is needed to develop (geo)visualization tools that can efficiently deal with all of the dimensions of Big Data, including

quality and veracity of the data. Ideally, the design of visualization should be informed by capabilities and constraints in human information processing, perception, and cognition (Li et al., 2016).

Many of the current challenges in visualization come from scalability issues. As the volume of data to be analyzed continues to increase, it becomes increasingly difficult to provide useful visual representations of data. In recent years, there have been advances in the visualization of data through various approaches with GIS-based capabilities. Better techniques and methods, however, are needed for analyzing Big Data, especially for massive and high-dimensional data sets that are heterogeneous in nature.

Interpretation is at the center of data analysis. Regardless of the size of the data, it is subject to limitations and bias. Without these biases and limitations being understood and outlined, misinterpretation tends to be the rule rather than an exception. Big Data is most effective when researchers take account of the complex methodological processes that underlie the analysis of data.

4 Cross-Cutting Challenges

Cross-cutting challenges are common challenges that underlie many, sometimes all, of the stages of the data analysis pipeline. Heterogeneity, uncertainty, scale, timeliness and human interaction problems with Big Data may impede progress at all stages of the pipeline.

4.1 Heterogeneity

When humans consume information, a great deal of heterogeneity is comfortably tolerated or even desired. However, machine algorithms expect homogeneous data, and cannot easily understand nuances. Consequently, one has to structure Big Data carefully as a first step in or before data analysis. To do this efficiently, one needs to express differences in data structures and semantics to be shown in forms that are computer understandable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution.

Unstructured data is difficult to work with, using relational database management systems and desktop statistics and visualization software. NoSQL (not only structured query languages) database management systems, instead, provide support for clouding architectures and the facility to generate patterns and trends without the need for additional infrastructure. Sometimes it is just not possible or practical to combine Big Data with varying spatial and temporal scales, hierarchies, and levels of resolution to make it compatible for analysis.

4.2 Uncertainty

Uncertainty is present in all stages of the Big Data pipeline. Representation and propagation of constrained forms of data quality, such as error bars, is an active area of research. There are several sources of errors in the process from data to inference and interpretation: errors and noise in the measurements themselves, lossy compression of data, mistakes in model assumptions, and (unknown) failures in algorithm creation and execution. Other sources of errors prevalent in Big Data include the high-dimensional nature of many data sets, issues of heterogeneity, and unknown provenance of data items in a data base.

If errors are present in the raw data, they can propagate to all stages in the Big Data pipeline. Recent work on managing probabilistic data and modeling suggests one way to make progress. For example, interval analysis allows one to model the uncertainty of the input variables (e.g., from sensor observations) and the corresponding uncertainty of the functions based on the variables (Li et al., 2016). Functional analysis methods (e.g., wavelets) are also useful for modeling uncertainty. Moreover, precision analysis can be used to evaluate the veracity of Big Data from the perspective of data quality, while simultaneously ensuring that the utility of the data is preserved.

4.3 Scale

In recent years, parallel and distributed computing systems have become a reality. These systems have given rise to search engines and online commerce and entertainment, providing the platform on which Big Data issues and problems came to bear. Scaling these systems and related algorithms to increasingly larger data sets is an ongoing challenge (National Research Council, 2013). Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processing getting faster, following Moore's law (Mishra et al., 2016). In recent years, there is a shift underway to move towards cloud computing. Cloud computing aggregates multiple disparate workloads with varying performance goals across large numbers of processors to manage computational efficiency. The most striking characteristic of cloud computing is its elasticity and ability to scale up and down as needed, making it suitable for data storing and processing in the Big Data era (Fan, Han and Liu, 2014). The level of sharing resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs and to deal with system failures. This task requires us to rethink how to design, build and operate data processing components to support activities at each stage in the data pipeline.

Given that Big Data can be highly dynamic, it is often infeasible to store and process the data in a centralized data base. The main approach to address this problem is to divide-and-conquer, which partitions a large problem into tractable and independent sub-problems. Each sub-problem is tackled simultaneously by different processing units (Fan, Han and Liu, 2014). Ensemble analysis, which strategically integrates multiple algorithms, can enable us to model an entire data set rather than a subsample of the data. Spatial ensemble methods may be applied to deal with the nuances of data. However, use of ensemble methods poses some difficulties, including ensuring that there is consistency between the algorithms. Moreover, many partitioning

techniques are not yet optimized for geometric computation (Wang, Aji and Vo, 2015). Another approach for managing scalability issues in Big Spatial Data is to exploit complex properties of such data, e.g., fractal patterns. Indeed, data produced via bottom-up mechanisms, such as crowd-sourced data, tends to exhibit fractal structure and related properties, which lends itself to such strategies (Batty et al., 2012; Li et al., 2016).

4.4 Timeliness

The design of a system that effectively deals with size is likely to result into a system that can process a given data set faster. However, it is not just this speed that is usually meant when one speaks of velocity in the context of Big Data. Rather, there is an acquisition rate challenge and a timeliness challenge. Many data sources operate in real-time, producing data streams that can overwhelm data analysis pipelines. And there is often a desire to make decisions rapidly, possibly also in real-time. If, for example, a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed, potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Instead, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination (Mishra et al., 2016).

For spatial algorithms, in particular, we cannot wait until all the data are known (Li et al., 2016). A significant requirement for data-intensive spatial applications is fast query response which requires a scalable architecture that can query spatial data on large-scale data. However, speed must not come at the sacrifice of the validity and trustworthiness of the data and results based on the data (Li et al., 2016). For useful large-scale, real-time analysis of Big Data, most if not all of the processes should be automated. While techniques like complex event processing and online analytical processing are useful for managing multiple, fast-moving data streams, they are not yet able to adequately support geospatial features and computations in an efficient manner (Lee and Kang, 2015).

4.5 Human Collaboration

Despite the progress achieved in developing machine-based solutions for processing and analyzing Big Data, humans still need to provide input into the data analysis loop at all stages of the pipeline (Mishra et al., 2016). While it may be able to find many 'statistically significant' results and effects with Big Data, discerning the substantive relevance and importance of these findings remains a challenge. However, drawing meaningful inferences tends to be subjective and context-dependent, and these are aspects of human intelligence that – at least currently – are beyond the capability of machines and algorithms. Thus, Big Data analysis must be evaluated with human subject matter knowledge and experience (National Research Council, 2013).

Humans are needed to understand the context, adequately frame analyses using Big Data, and position models in appropriate theoretical and empirical contexts. The new field of visual

analytics is attempting to do this, at least concerning the modeling and analysis stage in the pipeline (National Research Council, 2013). A popular new method of harnessing human ingenuity to solve problems is through crowdsourcing or participatory sensing, in situations where human perception, peoples' ability to disambiguate context and make subjective judgements, exceed the capabilities of computer systems (see Sui, Elwood and Goodchild, 2013, for crowd-sourcing of geographic knowledge). While collaborative mechanisms are a rich source of data, the data tend to be error-prone, biased, and context sensitive. Incentive-based approaches have the potential to help in improving the quality, cost and timeliness of crowd-sourced data. This is an important area for future research (National Research Council, 2013). In a smart city context, community participation and engagement are critical for ensuring the creation of reliable, timely and trustworthy information about collective phenomena (Batty et al., 2012).

5 Closing Remarks

The paper discusses computational and statistical aspects of the analysis of Big Data on cities and regions, using the data analysis pipeline as guiding framework. This final section summarizes some of the key conclusions.

First, Big Data on cities and regions are not centrally stored but distributed across multiple technical infrastructures, creating challenges in data access, integration, sharing and use. Accordingly, analysis of Big Data should make effective use of parallel and distributed hardware platforms, accommodating a wide range of data formats and statistical methods, and providing seamless interfaces to other computational platforms and tools for visualization of central aspects of the analysis.

Second, scalability is one of the most crucial technical challenges in Big Data analytics. Most of the methods available for analyzing data scale only to certain levels of complexity and size of Big Data. Beyond such levels these methods will become increasingly irrelevant and likely to be not appropriate for developing refinements of competitive value. Evidently, we face a need for new statistical thinking and computational methods to tackle the scalability challenges of Big Data. The design of computational procedures has to address challenges such as heterogeneity, noise accumulation and spurious correlations, and to balance statistical accuracy and computational efficiency.

Third, there are many sources of potential error in Big Data analysis, such as high-dimensionality and heterogeneity of Big Data, biases arising from sampling procedures and processes, and missing information on the provenance of the data. Any data analysis is based on a set of assumptions, and the assumptions underlying traditional statistical methods are not likely to be satisfied with Big Data. In addition, issues such as most notably that of sampling impinge on the quality of inference (National Research Council, 2013).

Fourth, Big Data analysis creates new challenges at the intersection between humans and machines. Human input from data analysts and domain experts is needed throughout the Big Data pipeline, specifically in instances where algorithmic approaches are insufficient in making sense of the data.

Finally, it is worth noting that research and development necessary to tackle the challenges of using Big Data on cities and regions does not involve a single discipline or field. To this point, there is a critical need for cross-fertilization among different subject domains. Not only do computer scientists, mathematicians and data analysts play an important role in this regard, but also experts in visualization including artists. Domain specialists and users of technology also have an essential role to play in designing and developing new perspectives on Big Data analysis and computation.

References

- Andrada, H.C.M., Gedik, B. and Turaga, D.S., 2014. *Fundamentals of Stream Processing. Application Design, Systems and Architectures*. Cambridge University Press.
- Batty, M., 2013. *The New Science of Cities*. MIT Press.
- Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G. and Portugali, Y., 2012. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1), pp.481-518.
- Bengio, Y., 2009. Learning deep architectures for AIT. *Foundations and Trends in Machine Learning*, 2(1), pp. 1-127.
- Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. Springer.
- Fan, J., Han, F. and Liu, H., 2014. Challenges of Big Data analysis. *National Science Review*, 1, pp.293-314.
- Fischer, M.M., 2015. Neural networks. A class of flexible non-linear models for regression and classification. *Handbook of Research Methods and Applications in Economic Geography*; Karlsson, C., Andersson, M., Norman, T., Eds., pp.172-192.
- Fischer, M.M. and Wang, J., 2011. *Spatial Data Analysis: Models, Methods and Techniques*. Springer, Heidelberg, Dordrecht, London, New York.
- Fischer, M.M., Scholten H., and Unwin, D. (Eds.), 1996. *Spatial Analytical Perspectives on GIS*. Taylor & Francis, Basingstoke.
- Gantz, J. and Reinsel, D., 2012. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East. *IDC iView: IDC Analyze the Future*, 2007(2012), pp.1-16.
- Getis, A., 1999. Some thoughts on the impact of large data sets on regional science. *The Annals of Regional Science*, 33(2), pp.145-150.
- Kourtit, K. and Nijkamp P., 2018. A big data dashboard architecture for a computable intelligent city. *Bollettino Del Centro Calza Bini*, 17(1), pp. 23-34.
- Kwan, M.P., 2016. Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, 106(2), pp.274-282.
- Lee, J.G. and Kang, M., 2015. Geospatial big data: Challenges and opportunities. *Big Data Research*, 2(2), pp.74-81.
- Li, S., Dragicevic, S., Castro, F.A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A. and Cheng, T., 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, pp.119-133.
- Manley, D., 2014. Scale, aggregation, and the modifiable areal unit problem. *Handbook of Regional Science*; Fischer, M.M, Nijkamp, P., Eds., pp. 1157-1171.

- Miller, J.D., 2016. *Big Data Visualization*. Packt Publishing.
- Mishra, B.S.P., Dehuri, S., Kim, E. and Wang, G.-N., (Eds.), 2016. *Techniques and Environments for Big Data. Parallel, Cloud, and Grid Computing*. Springer.
- National Research Council, 2013. *Frontiers in Massive data analysis*. National Academies Press.
- Osborne, J.W., 2013. *Best Practice in Data Cleaning*. Sage.
- Panda, M., Dehuri, S. and Patra, M.R., 2015. *Modern Approaches of Data Mining*. Alpha Science International.
- Pattnaik, K. and Mishra, B.S.P., 2016. Introduction to Big Data analysis. *Techniques and Environments for Big Data. Parallel, Cloud, and Grid Computing*; Mishra, B.S.P., Dehuri, S., Kim, E., Wang, G.-N., Eds., pp. 1-20.
- Raschka, S. and Mirjalili, V., 2017. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikrit-learn and TensorFlow*. Packt Publishing.
- Schintler, L.A. and Chen, Z. (Eds.), 2017. *Big Data for Regional Science*. Routledge
- Sui, D., Elwood, S. and Goodchild, M., 2013. *Crowdsourcing Geographic Knowledge, Volunteered Geographic Information (VGI) in Theory and Practice*. Springer.
- van der Maaten, L.J.P., Postma, E.O. and van den Herik, H.J. 2009. Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005. Tilburg University, The Netherlands.
- Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S. and Shekhar, S. (2012, November). Spatiotemporal data mining in the era of big spatial data: Algorithms and applications. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* (pp. 1-10). ACM.
- Wang, F., Aji, A. and Vo, H., 2015. High performance spatial queries for spatial big data: From medical imaging to GIS. *Sigspatial Special*, 6(3), pp.11-18.