Building streaming GIScience from context, theory, and intelligence

Carson J. Q. Farmer, Centre for GeoInformatics, University of St Andrews
Alexei Pozdnoukhov, National Centre for Geocomputation, National University of Ireland, Maynooth

Abstract

In this paper, we suggest that a focus on the current strengths of GIScience, coupled with a strategic view towards the future of data-driven research, can help to propel GIScience forward as a leader in data-intensive social science. With the majority of the world's data being embedded in space, it is only natural for GIScience to take a leadership role in the analysis and understanding of individual, local, regional, and global data by providing context, theory, and intelligence to an otherwise data-centric science. We suggest that in order to avoid the limitations of current data storage, management, and retrieval practices, a focus on real-time, intelligent analysis of data in a streaming framework is the most logical step forward. This type of analytical framework addresses two key concerns of GIScience: scalability and relevance. By focusing on results and models over raw data, we build on the current strengths of GIScience, leading to process-based research that is scalable over the long-run. Furthermore, by developing the methods and theories around streaming spatial data, we ensure that GIScience remains relevant in the increasingly data-intensive world of computational social science research.

Introduction

Data has always been big. Researchers, businesses, and government departments have continually collected and maintained large datasets relevant to their area of expertise. For decades, 'large' has been a moving target, chiefly dictated by the accelerating decrease in processing and storage costs. Despite a long history behind the use of large data sets for decision making and analysis by business and government, 'big data' has only recently emerged as an area of inquiry unto itself. This late emergence of big data is likely a function of several factors, including the fact that our ability to sense, collect, and process data from multiple sources is now far outpacing our ability to store and manage said data [1] (Figure 1). Additionally, we are beginning to see a major shift in the way many scientists are thinking about information and analysis, leading to a more data-intensive social science where hypotheses are generated through an abductive process (i.e., hypotheses are developed to account for observed data).

With increasingly efficient means of generating data from multiple sources, the amount and number of different types of data that industry and government collect on a regular basis has reached critical levels. Additionally, information pertaining to all facets of society, from public databases such as national census', to private customer databases, to community-built open data sources, are increasingly being linked to geographical locations. Indeed, as much as 80% of all information held by business and government may be geographically referenced [2,9], and this number is only likely to grow as more and more organisations realise the importance of locational information [6]. The massive amounts of data being collected, coupled with the additional complexity that spatial data yields, means that the

traditional GIS model of data storage and management is no longer sufficient, and that new insights into spatial data management and analysis are required.

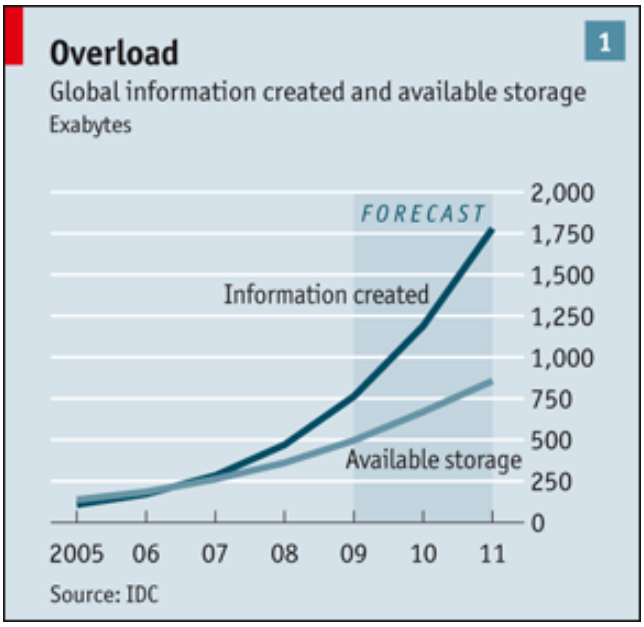

Figure 1: Global information created and available storage space. From The Economist (Feb 25 2012).

In the past, the limiting factor for geographic information science has been data. This is because GIScience has long been an abductive science, and inferring processes from patterns requires detailed information on both the location and attributes of the particular phenomenon under consideration. This data constraint is particularly relevant to GIScience due to the (previously) prohibitive costs associated with collecting data over large spatial extents. As such, despite rapid development of theory throughout the 1980s and 90s in GIS and quantitative geography, many new ideas were left untested due to a lack of tools and/or technology. Currently, GIScience (among other fields) is experiencing the opposite problem: the rapid pace of data collection is exceeding the reach of current theory and methods [8].

From a data management perspective, it is no longer feasible to store data using the the traditional geodatabase design. As GIScience continues to move from a set of tools and techniques for working with geographic data to a science of geographic information [4], our remit is shifting from the development of software and hardware solutions for capturing, storing, managing, retrieving, and disseminating geographic data to one of context, theories, and intelligence. Indeed, the true strengths of GIScience lie in its ability to provide substantive knowledge, develop geospatial thinking/literacy, ask the right questions, interpret data and outputs within the correct context, and understand the implications of research findings. With this in mind, the best way that GIScience can contribute to a data-intensive science is to focus on its strengths, and use its home field advantage in full when forced to enter a tournament with data-intensive computer science.

A way forward

While GIS development and design has, in the past, been done primarily within GIScience, it is unrealistic to assume that GIScientists should continue to develop software and hardware solutions for processing and disseminating geographic data on its own. These are things that computer scientists,

engineers, and software developers working on GIS and related information technologies can do, and do well. By focusing on higher-level problems and concentrating on what can be done with data rather than a focus on data itself, GIScientists can provide the context, theory, and intelligence required by other fields to provide solutions to spatial problems.

The theories and techniques behind spatial analysis, GISystems, spatial statistics, spatial data representation, and other developments within the realm of GIScience are being stretched to capacity by modern data sets. In a survey of fifty-eight key researchers in the field of spatial analysis (and more broadly, GIScience), overcoming methodological limitations imposed by large datasets was highlighted as a key challenge for the future [8]. Furthermore, with the ubiquitous adoption of web-based mapping systems and increasing awareness that 'space matters', GIScience as a field is now a net-exporter of methods and ideas [7], and maintaining the home field advantage in terms of spatial analysis will be a key deciding point in the success of GIScience in a data-driven world.

Context, theory, and intelligence

A GIScience focused on the implementation details of GISystems will develop 'users' rather than 'researchers', where users "will see the world through a lens defined by the constraints and principles of database design" [4] and legacy GIS thinking. Instead, we suggest that a more fundamental approach, based on using geographic knowledge and theories to refine our methods and expand our understanding of spatial processes is warranted. Here, the use of geographic information theory is useful, but not sufficient on its own. Process-based theories such as spatial interaction, spatial behaviour, and spatial diffusion should also be considered, and can contribute to a process- or model-based GIScience where models and linked information are used to solve fundamentally geographic problems. The key component to developing geographical intelligence is the integration of context (i.e., when, where, and what spatial patterns were generated) and theory (i.e., why and how do the observed patterns correspond to known processes).

For example, in studies of commuting, researchers are often interested in predicting the number of commuters traveling between a particular origin and destination pair, or the destination that a commuter at a particular origin might choose from a range of possible destinations. Here, the context is clear: the phenomenon under investigation is commuting and the spatial setting is (usually) some urban environment. As such, context encompases the problem scenario (i.e., predicting commuting flows), information requirements (i.e., counts, socio-economic factors, distances), required level of detail (i.e., macro vs micro), and target outputs (i.e., maps, model parameters, predictions). While context guides our treatment of the problem, theory provides the means to developing a solution. In the above case of predicting commuting flows, we can incorporate theories of spatial interaction (i.e., macro commuting) or spatial choice (i.e. micro commuting) to inform our models. This provides us with additional information requirements and a framework within which to compare our results (i.e., do these results make theoretical sense?). In this sense, developing geographical intelligence is a synergistic process: intelligence comes from understanding spatial processes, spatial processes can be approximated via models, models are directly informed by theory, and theory is inextricably linked to the context within which it operates.

With these points in mind, we suggest that the most logical step forward for GIScience is a model-centric view on analysis, where the focus is on real-time, intelligent analysis of data in a streaming framework. This type of research framework allows GIScientists to focus on the strengths of GIScience (i.e., ontologies, spatially-aware statistical methods and theories), and avoids problems associated with the storage and retrieval paradigm of traditional geodatabases. In the following section we describe the

benefits of streaming analytics, and outline a framework which would allow a model-centric GIScience to contribute to the big data agenda.

Focus on streaming

Applications where real-time analysis of millions of temporally varying spatially referenced samples over wide geographical areas are required are becoming an everyday necessity. In addition to increasing volumes of sensor data produced by city infrastructures, real-time data feeds of users' activities through various applications such as Twitter [12], Flickr, Foursquare and others are becoming increasingly available. Location-aware applications and location-based services have become popular in recent years, such that many data feeds now have a geographic element by default, thus becoming forms of volunteer geographic information. There is much potential for GIScience to explore spatial relationships in such data to understand spatial patterns emerging from low-level human actions and interactions.

Large data volumes and the complex mechanisms behind data generation processes require new analytical approaches which are flexible, non-parametric, computationally efficient, and able to provide interpretable results for modelling non-stationary and non-linear processes in data-rich situations. Machine learning offers a selection of online algorithms designed for streaming data, where it is assumed that every data sample can only be seen once and processed in constant time. Algorithmic solutions for such systems can be borrowed from the signal processing field, where streaming data have been studied for decades and efficient incremental methods for typical optimisation problems (e.g., least squares optimization, matrix inversion and decompositions) have been developed. Such approaches offer straightforward extensions of many spatial statistical models to be applied in real time to temporally-varying data streams.

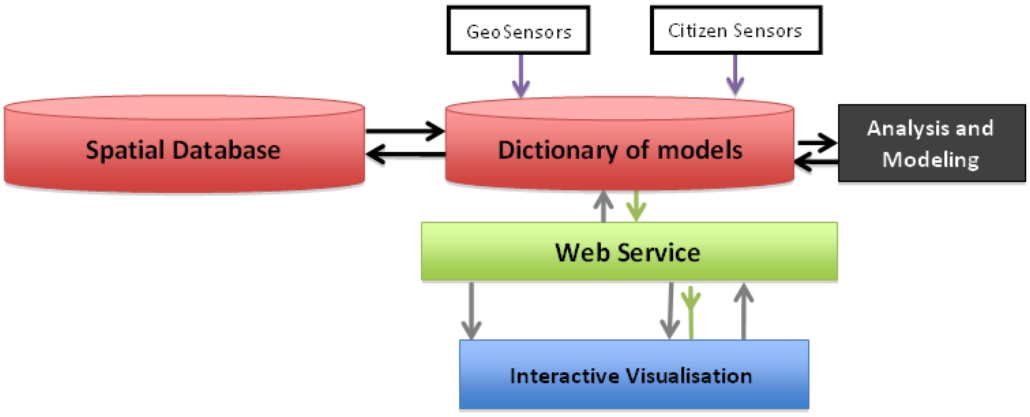

Figure 2: A transition from data- to model-centric design turns a GIS into a spatial knowledge discovery platform.

Excitement surrounding big data is continuing to build (see [1]), yet for the most part, big data analytics for spatial data has focused on data visualisation and descriptive analysis - a return to a data-intensive version of pre-1950s geography [5]. With a large range of spatial theories and domain specific techniques available, GIScience should not be dominated by simplistic description, but maintain its focus on theoretical understanding. A model-centric GIScience returns focus to modelling and understanding the underlying spatial processes rather than ongoing attempts to overcome the storage bottleneck. There is little sense in storing all the data if a model with a limited number of parameters is able to reproduce the phenomenon within a required level of accuracy: “perfection is achieved not when there

is nothing more to add, but when there is nothing left to take away[1]". This vision implies a different architecture for information systems which turns a GIS, traditionally centred around a spatial database, into a knowledge discovery system where we "bring the analysis to the data" and an analyst operates with a set of models (Figure 2) rather than with a raw data storage interface and a set of toolboxes. These models, expressed via a set of parameters and hyper-parameters, can be arranged in a dictionary trained on streaming data and capturing the typical (or atypical and thus interesting) state of the observed environment. Stream processing in this architecture is handled by incrementally updating model parameters [10], while particularly interesting samples can be detected and stored in a spatial database.

This focus on process over pattern is important for several reasons. Firstly, a model-centric framework improves interoperability of modelling spatially disparate datasets and simplifies data homogenization. For example, if an analyst is interested in understanding the relationship between temperature, humidity, and measured air pollution derived from different sensor networks, a spatially congruent set of data points is required. In this case, data interoperability is often achieved by querying the relevant databases and interpolating the values over a consistent spatial extent. In a model-centric framework, one gets this interoperability 'for free' by keeping up-to-date models of the various inputs. Similar advantages can be gained when dealing with multiple data streams attributed to spatial units of different types and geometries, and thus requiring polygon-to-point interpolation (see, e.g. [11]). Secondly, the decoupling of storage from modelling provides additional security and privacy by removing direct access to raw data. This means that potentially sensitive datasets may be analysed more freely/openly due to the 'anonymous' nature of models. An additional benefit of this type of analytical framework is its timeliness. Because inputs are analysed as they enter the system, results are continuously updated and therefore instantaneously relevant. The framework is also amenable to comparison, both spatially and temporally. Past models can be stored and retrieved when needed, providing the means to compare models through time, identify similarities in spatial dependencies, and provide quantitative decision support.

Conclusions

In this statement of interest, we have outlined how GIScience can take a leadership role in data-intensive social science research by focusing on its strengths rather than attempting to overcome the storage bottlenecks of legacy GIS systems. The focus here is on computational social science, as this is an area of research with a rich history of geographical theory that can be utilised to inform our models; however, a process-based framework may apply equally well in other GIScience related fields, including human-environment interactions, remote-sensing, and environmental monitoring. Indeed, by returning the focus to modelling and understanding underlying spatial processes, our framework for model-centric analysis provides a 'way forward' for GIScience research in general, that emphasises intelligent analysis, and provides models and tools to effectively explore, analyse, and understand dynamic spatial processes. This intelligent analysis goes a step further than the increasingly common practice of context-free data-driven knowledge discovery by taking advantage of contextual information and linked data to generate models that help us to represent and explain the real-world. Furthermore, a streaming GIScience helps to address concerns of scalability and relevance as we move towards a more data-intensive scientific paradigm. However, our focus on context, theory, and intelligence goes deeper than simply a better way to deal with large datasets; it may also help to address the long-standing concern of GIScientists that our lack of any unifying theories may reduce our field to second-class status in

[1] Antoine de Saint-Exupéry, Terre des Hommes (1939)

the academic community [3]. Changing the prevailing mindset, and focusing on modelling rather than storing, handling, and mining spatial databases will ultimately allow researchers to play to the current and future strengths of GIScience, and transform it into a leading discipline in the area of computational science.

References

[1] Cukier, K. (Feb 25 2010) Data, data everywhere. The Economist, Retreived from: http:// www.economist.com/node/15557443

[2] Franklin, C. (1992). An introduction to geographic information systems: Linking maps to databases. Database, 15(2) p12.

[3] Goochild, M. F. (2010) Twenty years of progress: GIScience in 2012. Journal of Spatial Information Science, 1:3-20.

[4] Goodchild, M. (2006) GIScience Ten Years After Ground Truth. Transactions in GIS, 10(1): 687-692.

[5] Golledge, R. G. (2008) Behavioral Geography and the Theoretical/Quantitative Revolution. Geographical Analysis, 40: 239-257.

[6]Hill, K. (Feb 5 2011) Internal Google Emails Shed Light On Importance Of Mobile Location Information. Forbes, Retreived from: http://www.forbes.com/sites/kashmirhill/2011/05/02/internal-google-emails-shed-light-on-importance-of-mobile-location-information/

[7] Longly, P. (2000) The academic success of GIS in geography: Problems and prospects. Journal of Geographical Systems, 2:37-42.

[8] Nelson, T. A. (2012) Trends in Spatial Statistics. The Professional Geographer, 64(1): 1-12.

[9] OGRIP. Advisory Committee's First Year Report. Columbus, OH: Department of Administrative Services, State of Ohio, 1990.

[10] Pozdnoukhov A., Kaiser C. Scalable Local Regression for Spatial Analytics, Proc. Of the 19th ACM SIGSPATIAL GIS'2011, 2011.

[11] Pozdnoukhov A., Kaiser C. Area-to-point Kernel Regression on Streaming Data, Geostreaming workshop at 19th ACM SIGSPATIAL GIS'2011.

[12] Pozdnoukhov A., Kaiser C. Space-Time Dynamics of Topics in Streaming Text, Location Based Social Networks workshop at 19th ACM SIGSPATIAL GIS'2011.