More bark than bytes? Reflections on 21+ years of geocomputation

Richard Harris*, David O'Sullivan, Mark Gahegan, Martin Charlton, Lex Comber, Paul Longley, Chris Brunsdon, Nick Malleson, Alison Heppenstall, Alex Singleton, Daniel Arribas-Bel, Andy Evans

*contact author:
School of Geographical Sciences, University of Bristol, rich.harris@bris.ac.uk

Abstract

This year marks the 21st anniversary of the International GeoComputation Conference Series. To celebrate the occasion, Environment and Planning B invited some members of the geocomputational community to reflect on its achievements, some of the unrealised potential, and to identify some of the on-going challenges.

Key words: geocomputation, urban analytics, Big Data, agent-based modelling, quantitative geography

---

Introduction

2017 marks the 21st anniversary and homecoming of the International GeoComputation Conference Series, started in Leeds in September 1996. The Nintendo 64 was released the same year. Two decades later, that company's most recent console is described as a hybrid, merging the handheld and home gaming experiences. Geocomputation also is a hybrid, fusing together the geographical and the computational. Has 21 years of development created something original and innovative or is it an idiosyncratic outsider searching for mainstream acceptance?

To celebrate the occasion and as part of Environment and Planning B's refocusing on urban analytics and city science – both areas of geographical and computational interest – we invited eleven well-respected members of the geocomputational community to reflect on some of its achievements, some of the unrealised potential, and some of the on-going challenges in the age of 'Big Data'.

What exactly is geocomputation if not an excessively syllabic portmanteau? As David O'Sullivan observes (below), the geocomputation community has struggled to forge a distinct answer and identity beyond "doing geography with computers." In the fourth edition of the Dictionary of Human Geography (Johnston et al., 2000) it is described (by Paul Longley, 2000) as the creative and experimental application of geographic information technologies in research that emphasises process over form, dynamics over statics, and interaction over passive response. Its appearance in the Dictionary, just four years after the first conference, suggests an early degree of academic credibility – of it doing something geographical that is not only recognisable but distinctive.

To gauge the success of geocomputation, Mark Gahegan looks back to that first conference, and to the first paper, recalling the eight challenges it presented. He notes their common theme, "to compute our way to better analytic solutions to geographical problems." In this regard, we may regard

geocomputation as prescient – an early response to the rising tide of ever more powerful computing and to the deluge of data it washes upon the shores of geographical interest because so much of those data are georeferenced. The optimist may see in this the opportunity to reinvigorate the ambition of spatial science: to further our understandings of spatial interactions and of spatial processes – time-space geographies, for example, the interactions between people and places, between urban forms and functions, about how cities evolve or 'work' as (chaotic) systems, or about how people behave and make decisions in different spatial places and contexts, and under varying social, economic and other constraints. And, it may do so in a way that lets the data do the talking through the brute force of computing; or, at least, in a way that is as much interested in exploring data geographically – searching for spatial variation, looking for localised departures from a general trend, finding something new, unexplained and spatially clustered – as it is about trying to 'prove' (in a statistical or classically econometric sense) more generalised 'laws' and theories.

However, rising tides deposit rubbish. As both Paul Longley and Lex Comber are aware, the freedom of geocomputation needs to be balanced against producing practical and usable findings that have at least some anchoring in theory, testable propositions and realistic representations of the observable geographic world. In addition, users should be suitably critical of what the data have and have not measured, and of the results they generate. The well-worn maxim of garbage in, garbage out still applies. However, data deluge need not lead to data junk if suitable checks and balances are in place, including what Chris Brunsdon advocates as reproducible research. The suite of localised and geographically weighted statistics outlined by Martin Charlton epitomise the coupling of the geo and the computational, grounded within a statistical framework to search for and not ignore the geographical patterning of a variable across a map. At a minimum, such methods provide a diagnostic tool to check the assumption of independence that infuse most statistical methods, including regression. But more than that, they challenge the whole idea of 'averaging away' spatial differences on the not unreasonable basis that those differences, and the processes that caused them, ought to be of some geographical interest.

If a goal of geocomputation is indeed to model social and economic processes, then on face value agent based models tick all the right boxes as they use data, computation, simulation, rules and randomisation to explores the links between theory, processes and geographical outcomes. Nick Malleson is hopeful that with the sorts of data collected under the rubric of smart cities, geocomputation has the potential to create reliable forecasts of urban dynamics. Alison Heppenstall is more questioning of the current state of play and its ability to model how real-world individuals really behave.

Therein lies the challenge. To quote Alison, "how can we use new forms of data to understand how real people shape and are shaped by geographical processes?" Phrased more broadly, how does all this computational power and all these data get us beyond measuring spatially differentiated outcomes to understanding better the processes that created those outcomes in the first place? How do we validate what we think we know about those processes and on what basis do we develop or discount existing theories? How does geocomputation engage with and contribute to the best of quantitative social science? And how do we do this in a way that has a wider impact, not locked away in the ivory towers of academia but engaging with commercial partners and teaching students the geocomputation skills they need to contribute to what Alex Singleton and Daniel Arribas-Bel call Geographic Data Science?

Looking back, it is clear that geocomputation has inspired a lot of computational and methodological innovation. Nevertheless, 21 is a coming of age. Apparently, the Switch is the fastest-selling console in Nintendo history. Can geocomputation also shape something distinctive in an era of knowing more

yet understanding less (Lynch, 2016)? Andy Evans is optimistic. If it holds to what he describes as the core principles of rigor, sympathy, and imagination, geocomputation will continue to inspire, innovate and evolve, and there will be plenty more celebrations ahead.

Richard Harris
School of Geographical Sciences, University of Bristol

What Geocomputation is *For*: Doing *Geography* with Computers

The geocomputation community has struggled to define itself clearly, and often is perceived as a quirky offshoot of geographical information science (GISci). However, self-consciously emerging in 1996 at the inaugural Leeds conference, a few years after Goodchild's (1992) calling into existence of GISci, it was clearly intended *not* to be GISci. Gahegan (1999) forcefully distances geocomputation from the "disabling" technology GIS, which has itself distanced quantitative geographers from geographical questions:

> Geocomputation is a conscious attempt to move the research agenda back to geographical analysis, with or without GIS in tow [...] It is about not compromising the geography, nor enforcing the use of unhelpful or simplistic representations (p. 204).

It is difficult to argue plausibly that this goal has been achieved. Yet I am more optimistic now that it *might* be achieved, than I have been for some time.

Mark Gahegan's pithy argument bears revisiting. In essence, he suggests that GIS has been a "Disabling Technology" (op. cit., p. 203), because "GIS saw to it that geographers became the slaves of the computer, having to adopt the impoverished representational and analysis capabilities that GIS provided." Of course, there are advantages to adopting shrink-wrapped computer solutions, among them "getting some sleep and producing much prettier output" although anyone who has attempted to bend an obstinate GIS package to their will, might quibble with even this modest claim.

More substantively, a side-effect of the widespread adoption of GIS in government, business, education and beyond, has been the actualization of early GIS-boosters' dubious (at the time) claim that 80 per cent of all data has a spatial component. The source of this often-cited boast is unclear. The earliest I have managed to trace it is to a conference paper by Antenucci (1989) but the context makes clear that it was by then already a commonplace assertion. Chrisman (pers. comm.) suggests it was routinely made to persuade doubting local government purchasing officers of the wisdom of investing in then untested GIS software with uncertain utility. In any case, 80% seems a likely underestimate now looked at 30 or 40 years later. Data today are routinely encoded with a spatial reference at the moment of collection, be it an address or GPS coordinates, and if not can be readily associated with a spatial location in a matter of minutes. This is thanks to the astonishing success and widespread adoption of GIS and more recently web-mapping and related technologies.

Nevertheless, the "simplistic representations" which Gahegan bemoans remain. For the most part, the geography associated with data is encoded as a point, or a polygon. Together with other points or polygons these are assembled in spatial layers. Notwithstanding the many operations that contemporary GIS software can perform on and between spatial layers – which the web developer community at the time of writing is assiduously reinventing – the limits of points, polygons (and also grids) as representations of geography are apparent. Geographers, as a rule, are more interested in the (spatial) relations among things than in the things in themselves and in how processes play out at

multiple scales. How do spatial relations affect how things change over time, and how do those relations change over time and across scales as a result?

Such *geographical* concerns lay at the heart of geography's quantitative revolution of the 1960s and persisted into the 1970s (see Forer 1978, for example). It was only as the initial hope of incorporating space into statistical tools proved trickier than expected (Gould 1970), and the hoped-for one-to-one correspondence between processes and the patterns they produce proved a mirage (Olsson 1969) that the confidence (hubris?) of quantifiers waned. Meanwhile geography embraced other epistemologies, and some of the creative energy of would-be quantifiers was directed into building mainstream GIS and its accompanying infrastructure of data models, ontologies, algorithms and routinized analytical approaches.

Somewhat in the shadow of these developments we have seen the emergence of more open-ended, platform and data-model agnostic tools for the analysis of geographic data (see Brunsdon and Singleton 2014). This alternative geospatial ecosystem now seems ready for widespread adoption by geographers, without the same commitment to particular approaches to representation that GIS demands and subtly enforces. Geocomputation seems an apt label for this polyglot assortment of approaches. After all, as Helen Couclelis (1998) noted, if it weren't for the happy accident of the pronounceability of 'geo' as a prefix, we'd likely call it 'geographical computation'. And what else would a geographer with a computer be interested in?

David O'Sullivan
Department of Geography, University of California, Berkeley

Geocomputation's 21 year report card: B-, some good progress, but could try harder

Geocomputation began in earnest with the conference at Leeds University in 1996 and rapidly became established as a vibrant research community (papers from the first gathering and all subsequent meetings are available at www.geocomputation.org). In the very first paper describing this new field, Stan Openshaw and Robert Abrahart (1996) defined a series of eight challenges that, for them, defined the direction (here paraphrased for brevity):

1. improving the resolution and precision of computational models;
2. computationally intensive statistical methods such as jack knifing and bootstrapping or the use of Monte Carlo significance tests in place of heavily assumption dependent classical alternatives;
3. improved optimisation methods that can use stochastic search or evolution strategies;
4. unsupervised learning methods to replace simplified statistical tools;
5. improving supervised computational models by removing simplifying assumptions, via neural methods;
6. adding more geographical knowledge into a problem (for example by using fuzzy logic);
7. tools for data mining, pattern recognition and cluster detection, including artificial life methods, that can search large data spaces;
8. application of new search techniques from machine vision.

Of these, the first three aim to leverage improvements in computational speed and scientific computing to offer more accurate, more scalable analysis, or the use of previously intractable statistical methods. The remainder aim to leverage what were recently pioneered techniques in machine learning and artificial intelligence. The common theme linking these challenges is the desire to compute our way to better analytic solutions to geographical problems, by continuously improving methods and leveraging Moore's Law. Geocomputation is an apt name for such a field.

How far have we got with this original agenda? Let's look at the high computing challenges first. At the beginnings of geocomputation, there was significant interest in high performance computing and GIS (e.g. Armstrong, 1995; Healey et al, 1997). At the time (1996), many universities had access to gigaFLOPS computing platforms—that is 10^9 floating-point operations per second, and the world's fastest computer could manage around 1 teraFLOPS (10^{12} operations/sec). Access to this kind of computing power opened possibilities for many research communities in terms of new analysis methods and scaling up longstanding problems such as global climate and ocean circulation modelling.

Two decades on and many single CPUs can now sustain over 1 teraFLOPS; some researchers have access to petaFLOPS (10^{15}) machines. I doubt there is an analytical question currently posed in geography that would need more compute power than that of the world's fastest computer (around 100 PFLOPS). But the problem is not raw power, it is scaling up our algorithms so that they can take advantage of such platforms via parallelization and optimisation. In this regard, geocomputation has achieved very little in the last twenty years: the re-expression of spatial algorithms and data structures onto established HPC templates (Asanovic, 2006) has proceeded intermittently with little concerted effort, a notable exception being the work to parallelise the GRASS open-source GIS (Akhter et al., 2010). However, there has been a late resurgence of interest in this topic, in large part due to the overlap of goals with CyberGIS and related cloud computing initiatives, (e.g. Shi et al., 2013; Satish, 2015; Stojanovic and Stojanovic, 2013).

Turning to machine learning and artificial intelligence, the report card is better. Machine learning techniques such as neural networks, decision trees, genetic algorithms and artificial life have received a steady stream of interest. Papers experimenting with their application in spatial analysis, remote sensing, and ecology appear quite regularly in the literature (e.g. Fisher, 2006; Wiley et al, 2003; Pijanowski, et al., 2002; Gahegan, 2000). More recently, the focus of such papers has moved from explaining and justifying these new methods to getting the best out of them and demonstrating how much better they are than simple statistical approaches (Rogan et al, 2008; Pradhan, 2013). Related interest in geographic knowledge discovery (Miller and Han, 2009) has also helped to further this part of the geocomputation agenda.

Despite their clear improvements in predictive power, machine learning methods remain notably absent from commercial GIS and remote sensing software. The challenge in moving them towards mainstream adoption is twofold: machine learning methods usually require experimentation with various configuration and learning parameters to get the best out of them, which makes them difficult and time-consuming to use, especially for non-experts; (ii) the statistical models that machine learning challenges are often simpler to apply, more stable (results do not vary due to search heuristics) and the error or goodness of fit is computable. However, the first of the challenge may already have been overcome. Deep learning methods—often based on hierarchies of neural networks—are proving to be effective at many learning tasks, as they essentially remove or streamline much of the difficult setup and experimentation phase; essentially this too is solved by the network as part of its learning process (Yann et al, 2015; Schmidhuber, 2015).

Moving from machine learning to the reasoning and automating aspects of artificial intelligence (AI), the progress is slower. A review of the possibilities is provided by Wu and Silva (2010) and some of the practical benefits and challenges are discussed by Malerba et al, (2003).

On reflection, I believe the biggest contribution that geocomputation has made in the last twenty years is to encourage a generation of scholars to experiment with new computational methods and with their application to geographical problems. Given that geographers do not always have a strong background in computer science, some of these methods can be challenging to understand and difficult to apply. It is immensely rewarding to see that so many researchers have tried, succeeded, and made geographical analysis and modelling richer and more powerful as a result.

Mark Gahegan
Centre for eResearch and Deprtment of Computer Science, The University of Auckland, New Zealand.

Geocomputation: a geographically weighted success story

In the early 1980s there was interest in the association of cancer 'clusters' with supposed sources of radiation contamination. The media also was concerned with possible contamination arising from the nuclear waste reprocessing plant at Sellafield on the Cumbrian coast. The Black Advisory Committee (Black, 1984) concluded that the Sellafield plant was not connected with raised levels of leukaemia in Cumbria, but recommended re-analysis of local cancer registries. Stan Openshaw and colleagues at Newcastle University undertook the analysis, leading to his seminal paper (Openshaw et al, 1987) which appeared in the first volume of the fledgling *International Journal of Geographical Information Systems*.

Existing approaches had identified a source of radiation in the electromagnetic spectrum to determine whether the rate of morbidity around it was somehow higher than some national level. My recollection is that electricity substations were regarded as suspicious, as were electricity powerlines. But could the sources include telephone boxes or fish and chip shops.? The underlying issue was that no-one knew what the linkage might be.

Stan inverted the problem and decided that if he could determine where the excesses were centred, this might lead to more fruitful line of enquiry. Thus a *whole-map* statistic that suggested evidence *of* clustering was replaced by a *local* statistic that suggested *where* that clustering was located. The implementation led to a range of computational and statistical challenges but those do not diminish the importance of Openshaw et al (1987) and subsequent papers.

That was 30 years ago. We can see other stimuli to geographically-minded approaches. In the early 1970s, Emilio Casetti (1972) had conceived of regression parameters that might exhibit heterogeneity; his ideas were subsequently extended by John Paul Jones III (Jones and Casetti, 1993). Wilpen Gorr had experimented with parameter 'drift' in regression models at around the same time (Gorr and Olligschlaeger, 1994), and Luc Anselin had looked at both modelling spatial structure and local statistics, in particular local indicators of spatial association (Anselin,1995). Rogerson (1999) developed a local chi-square statistic to examine evidence for disease clustering in New York.

Work at Newcastle University in the early 1990s lead to the first paper on geographically weighted regression (Brunsdon et al, 1996). A subsequent book by the same authors (Fotheringham *et al*, 2002) consolidated their previous work and presented new material. A paper on GWR was also presented at the first geocomputation meeting at the University of Leeds. However, an early issue was software. We forget that Openshaw's GAM code was written for the IBM and Amdahl computers at Newcastle

University running a unique operating system developed at the University of Michigan. In the mid-1980s rapid data exchange involved a 12 inch diamater reel of magnetic tape and a courier service. Embryonic FORTRAN code for GWR was made available for potential users to download. Windows software for GWR was later available at cost from Newcastle University following the Fotheringham et al (2002) book launch. A more advanced Windows application has been available in the last few years.

The award of a Strategic Research Centre to the National University of Ireland Maynooth by Science Foundation Ireland provided an opportunity to develop the geographical weighting approach. The major output was a package of open source code for the R system: GWmodel (Lu *et al*, 2014; Gollini *et al*, 2015). This extends the previous developments considerably, and includes functions for univariate and bivariate analysis, generalised linear models, ridge regression, discriminant analysis and principal component analysis. Criticism of the susceptibility of GWR to collinearity among the predictor variables has been addressed by the development of locally compensated ridge regression (Gollini *et al*, 2015). The package also includes functions to allow the use of different distance metrics in the geographical weighting, including network distances and the Minkowski metrics.

Chris Brunsdon has observed that the Pearson correlation coefficient can be unpicked as a LISA. If the two variables x and y have been mean centred, then the values:

$$r_i = \frac{x_i y_i}{\sqrt{\sum_{j=1}^{N} x_j^2 \sum_{j=1}^{N} y_j^2}}$$

are the individual components that sum to the value of *r*. These can be plotted on a map to show which locations contribute the most, or least, to the value of the correlation coefficient. Such an approach complements the geographically weighted correlation functions in GWmodel. Recent work includes the development of geographically weighted correspondence matrices (Comber *et al*, 2016). GWR appears as a tool in the Spatial Statistics toolbox of ESRI's ArcGIS software, and for users of Quantum GIS, it is available in the SAGA freeware that installs alongside QGIS. There are versions in other packages (including spgwr and gwrr).

At the time of writing (early 2017) the search string *geographically weighted regression* returns 79600 hits in Google. It's a fitting tribute to a public health scare, a mainframe 'super computer' and a visionary academic.

Martin Charlton
National Centre for Geocomputation, National University of Ireland Maynooth

Geocomputational Musings on Big Data

There is a great deal excitement across many scientific communities about the new opportunities afforded by Big Data. For the geocomputation community, the potential lies in Big *Spatial* Data, and the opportunities to harness the increasing number of open data initiatives, new forms of data generated by citizens, the near ubiquitous capture of location, and the near permanent connectivity via web-enabled devices that allow data to be shared and uploaded.

Classically research is undertaken in the following way:

1. Formulate a research question.

2. Identify what data to collect and how to collect it.
3. Perform some statistical tests to determine whether any effects or associations arise due to random sampling errors.
4. Get an answer to the question.

Big data turns experimental design and its associated inferential theory on its head:

1. Collect lots of data about anything.
2. Perform some kind of data mining.
3. Get some kind of answer.
4. Decide what question it was an answer to.

A common theme is to allow the data to do their own talking, with the potential for data mining and machine learning to identify important but hidden associations of social or scientific interest:

> Scouring databases and other data stores for insight is often compared to the proverbial search for a needle in a haystack, but ... big data turns that idea on its head … [and quoting Viktor Mayer-Schönberger] "With big data, we don't know what the needle is. We can let the data speak and use it to generate really intriguing questions" (Needle, 2015).

The idea is attractive but also empirically and theoretically naive. If research questions are not specified in at least some sense in advance, then the results of data mining risk being answers to arbitrary questions. If the aim is to find a needle in a haystack then making the haystack bigger does not make the job any easier, and if we don't know what kind of needle we are looking for, it helps even less. In the shadows of the Big Data paradigm is a need to revisit classic tools for statistical inference (Brunsdon, 2016). This is because of the ease with which spurious, nonsensical relationships and correlations between variables can be inferred through data mining, and because of the lack of rigorous statistical methods for analysing very large datasets, where statistical 'significance' is meaningless.

Paul Mather and Stan Openshaw summarised these concerns in a prescient way in 1974. Reflecting on the potential to analyse population census data using computers they suggested:

> It might be far more profitable to postulate a certain pattern of factor loadings (and inter-factor correlations) and attempt to find **how far the hypothesis fits the data** that has been collected. This attitude should help prevent the **mindless approach** in which numbers of variables characterized only by the fact that they are all easily culled from census volumes or derived from two or three basic variables, **are picked over like cans on a rubbish tip**. (Mather and Openshaw, 1974, p290, emphasis added).

Geocomputation can play an important role in addressing these concerns. The process of big data analysis should be a process of investigation driven and supported by some sort of theoretical underpinning. Where these are absent, then analyses should proceed with reflective cycles of investigation and explanation, rather than simply data mining and hypothesis testing – it should be explorative detective work perhaps aided by visualisation. The importance of exploratory analyses cannot be overstated: they support the iterative development of theoretical constructs as the basis for analysis *and* to develop robust and reproducible big data analyses by looking for patterns (geographical and otherwise) through repeated experiment. For example, in the absence of theory this could be by randomly sampling the big data, identifying patterns, applying to other samples or to

the whole dataset, and then engaging with domain experts to anchor the results in a theoretical framework for the study.

In short, geocomputational analysis should be grounded in some idea of what questions are important. The reflexive process described above supports that identification. Big Data analyses should include a reflexive cycle of investigation and explanation, rather than data mining, repeat testing (exploring rather than fishing) and it should support the iterative development of theoretical constructs as the basis for analysis. Until we act in this way Big Data analyses will not help us to answer the Big Questions we currently have, nor identify new Big Questions deep in the Big Data.

Lex Comber
School of Geography, University of Leeds

Big Data, Geocomputation and Geography

Unlike some other areas of computer intensive programming, geocomputation has fallen somewhat short in delivering transparent models with practical, usable findings. This is perhaps disappointing in contrast to (a) the obvious application success in embodying core principles of spatial organisation in geographic information systems and (b) the vast streams of spatially and temporally referenced data that have become available in many applications areas.

In computational terms, it seems likely that this is because the geo-temporal frame that are subject to analysis are unbounded. 'Geographic' is commonly taken to imply scales from the architectural to the global and work presented at the very first geocomputation conference in 1996 illustrated that issues of scale and recursion opened up a seemingly infinite range of ways of framing representation of the world – in terms not only of form and process, but also statics and dynamics. Any representation of how the world looks and how it works is therefore necessarily partial, incomplete and, in temporal terms at least, open-ended. Contrast this with the closed system computational problems of, say, translating natural languages (where the system is bounded by finite dictionaries of words and grammatical structures), and it is perhaps unsurprising that the achievements of geocomputation are more muted. Piecemeal and partial models achieve piecemeal and partial outcomes and there is some inevitability that this wil be the case.

The 'geo-' prefix differs from its 'spatial' counterpart not just in the range of scales that it may describe, but also in its implied association with the unique place of the surface and near surface of Planet Earth. This fundamental distinction may have underpinned some of the 'GIS wars' between Openshaw and Taylor, and others in the 1990s (Openshaw, 1991; Schuurman, 2000) – analysis of the canals of Mars may meet the scale range criterion but does not fulfil the place criterion and as such, sensu stricto, does not qualify as geographic analysis. This distinction highlights that geography as a discipline brings tacit knowledge to understanding of places that are fundamentally unique accretions of the outcomes of past human and social processes. Representations of place need to provide an effective base for geocomputational analysis of the general effects of current and future geocomputational processes. This is because geographic objects of analysis are not simply locations in space but the accumulated outcomes of systems of networks and flows (see Batty 2013).

Understanding what 'place' is or at least how it can be effectively represented presents daunting application specific challenges. Clear conceptions of the nature of the geographic data are required, yet there must be some concern that geocomputation has acceded to wider tendencies in Big Data analysis to disregard the provenance and quality of the huge volume and variety of data that are

available today. A generation ago students of social and environmental science were, it seems, much better versed in widely accepted scientific principles of research design, as well as the statistical apparatus of generalisation. This is not to say that there are never instances where the availability of billions of data points and ever greater data content cannot be a substitute for some vagaries of geographic coverage – sometimes there is a trade-off between the largely unknown biases of unconventional data but the spatial and temporal precisions that they can bring. But precision is not the same as accuracy, and representations of the world need to be accurate it they are not to prove biased, partial and potentially delusional.

How might geocomputation better respond to the challenges of a world in which is data rich but in which new forms of data do not provide anything that might be described as spatial data infrastructure? A first way is to better use what we know from conventional data sources that may be less detailed or up-to-date but which are of known provenance in terms of content and coverage. Machine-learning methods, for example, must be guided by clearly defined populations of interest. Geographical heuristics may be used to achieve the same ends. The richness and variety of Big Data make it possible to ground many more assumptions at highly disaggregate scales and, suitably triangulated with conventional framework data sources, draw inferences that are both robust and open to scrutiny. Current research using consumer data, which account for an increasing real share of all of the data collected about citizens, provides one relevant application area in this context (e.g. see cdrc.ac.uk).

Paul Longley
Department of Geography, University College London

Reproducible Research, Quantitative Geography and Geocomputation

A large proportion of practical quantitative work in geography relies on the analysis of data or on the running of simulation models. That analysis, and the results it generates, are the outcome of a process involving data verification, re-formatting, computer programming, modelling, data analysis and visualization. Many publications are created to share the results and to discuss their implications. It follows that the validity of the publications, and of future publications citing them depends on the validity of the initial work and on the analytical process. However, although the publication itself is widely available (in many current situations it is open access), details of the supporting activities - in particular, the data collected, the software code used and the exact stages of analysis - are often not available, or at least not easily traceable.

The term *reproducible research* (Claerbout 1992) is used to describe an approach which may be used to address this problem, and allow code and data to be easily accessed. Although not noted greatly in quantitative geography at the time of writing (but see Brunsdon and Singleton 2015) it has gained attention in a number of applied fields where quantitative data analysis is used, exemplified here by statistics (Buckheit and Donoho 1995; Gentleman and Temple Lang 2004), econometrics (Koenker 1996) and signal processing (Barni et al. 2007). The ultimate goal of reproducible research is that complete details of any reported results and the computation used to obtain them should be freely available, so that others following the same procedures and using the same data can obtain identical results. This approach is offered, for example, when using *Rmarkdown* - where data analysis code written in the R statistical programming language is incorporated into a text document. On viewing, the code is run and the output (either textual or graphical) is substituted into the document.

Distributing documents in this way, together with sufficient information to access the data analysed facilitates an open and reproducible approach to data analysis and visualisation.

A strong case can be made for a focus on this topic in quantitative geography, geocomputation and spatial science. The practice allows others to scrutinise not only the data used as the basis for an analysis, but also the approach to the analysis itself, creating a platform for greater scrutiny and accountability. A large amount of work involving the analysis of geospatial data influences policy in many fields - health, climate change and crime prevention are a small but significant set of examples.

The key justification of a reproducible approach is precisely that: it can be reproduced and validated by others. However, there are additional benefits: Reproducible analyses can be compared: different approaches addressing the same hypothesis can be compared on the same data set, to assess the robustness of any conclusions drawn. Also, methods used are portable: code can be obtained from documents, allowing others to learn from other people, to apply the code to other data sets and to adapt the code for related problems. Finally, results may be updated in situations where updated versions of data are published (for example new census data) and methods applied to to the original data may be re-applied.

Thus there are several arguments for reproducibility in quantitative analysis of spatial data - not just for academics, and not just for the geocomputationally minded, but also for public agencies and private consultancies charged with analysing data that may influence policy. Recent work (Vandewalle, Kovačević, and Vetterli 2009) has shown that papers in a number of fields adopting reproducible approaches have higher impact and visibility.

Achieving reproducibility like this is clearly within reach in some situations, although there are also some challenges ahead, as the diversity, frequency and volume of geographically information increases. Even in situations where personal or sensitive information is analysed it could be argued that there are advantages to having 'domains of reproducibility' – that is, groups of people who are permitted to access this information adopting reproducible practices amongst themselves – so that internal scrutiny, and updating of analyses becomes easier. Adopting reproducibility calls for some changes in the practice of both analysts - in adopting reproducible practices, and learning new skills, and publishers - who in support of this would need to provide resources where reproducible document formats may be submitted, handled, distributed and viewed by a wide audience. However, such changes are already taking place in other disciplines – for example in the journal *Biostatistics* – so why not in the field of geocomputation?

Chris Brunsdon
National Centre for Geocomputation, National University of Ireland Maynooth

Big Data, Agent-Based Modelling, and Smart Cities: A Triumvirate to Rival Rome

Following the Big Data revolution (Mayer-Schönberger and Cukier, 2013), aspects of peoples' lives that have never before been documented are being captured and analysed through our use of smart-phone applications, social media contributions (Croitoru et al., 2013; Malleson and Andresen, 2015), public transport smart cards (Batty et al., 2013), mobile telephone activity (Diao et al., 2016), debit card transactions, web browsing history, and so forth. Taken together, and supplemented with knowledge about the physical environment (air quality, temperature, noise, etc.), pedestrian footfall or vehicle counters (Bond and Kanaan, 2015), these data provide a wealth of *current* information about the world, especially cities. This "data deluge" (Kitchin, 2013) has spawned interest in 'smart

cities'; a term that refers to cities that "are increasingly composed of and monitored by pervasive and ubiquitous computing" (Kitchin, 2014).

One aspect to smart cities, largely absent in the published literature, is the ability to forecast as well as to react. Whilst most initiatives inject real-time data, these data rarely are used to make real-time *predictions* about the future. Where 'forecasting' is an advertised capability of a smart city initiative, rarely is it explained in any detail. This might be due to the proprietary nature of many initiatives but it is equally likely that a lack of appropriate methods is at fault. Although 'black box' artificial intelligence methods are progressing rapidly, there is little evidence that these are being used to forecast future states of smart cities.

Perhaps agent-based modelling offers the missing component for predictive smart cities? Agent-based models (ABMs) simulate the behaviour of the individual components that drive system behaviour, so are ideally suited to modelling cities. A drawback with ABMs is that they require high-resolution, individual-level data to allow reliable calibration and validation, and traditionally these have been hard to come by. However, in the age of the smart city, this no longer is the case. Furthermore, ABMs are not 'black boxes'; the individual agents are imbued with behavioural frameworks that are (usually) based on sound behavioural theories. This makes it easier to dissect the models, as well as allowing a controller to manipulate the behaviour of the agents as required for a particular forecast. In addition, because many 'big' data sources are available in *real time*, there is the opportunity to calibrate models as soon as new data become available. This is akin to forecasting in fields such as meteorology, where the latest weather data are assimilated into running models to improve short-term predictions. This triumvirate of big data, agent-based modelling, and dynamic calibration has the potential to become the de facto tool for understanding and modelling urban systems.

There are, however, substantial methodological challenges that must be met, including developing the means to assimilate the data into models. Furthermore, engagement with smart devices is not heterogeneous across the population, so there is a risk that those individuals who choose not to use 'smart' technology will be forgotten about in simulations and planning processes. Simulations that are based on biased data have the potential to increase biases by presenting biased results that are then used to influence policy. For example, PredPol is an extremely popular predictive policing tool that is being purchased by police forces across the globe in order to predict where future crimes are going to take place. However, policing data are biased towards particular minorities as a result of where most policing activity already takes place, so the tool has the potential to increase those biases by sending more officers to areas that are already being heavily policed (Lum and Isaac, 2016). Any smart city modelling/forecasting tool must be able to mitigate against these risks.

To conclude, although smart city initiatives are numerous, very few can evidence an ability to create reliable forecasts of future city states. However, advances in spatial methods that fall under the umbrella of 'geocomputation' have the potential to create reliable forecasts of urban dynamics under a variety of conditions. There are ethical issues that must be considered but, if conducted safely, the triumvirate of agent-based modelling, big data, and dynamic calibration is extremely attractive.

Nick Malleson
School of Geography, University of Leeds

ABM and Geocomputation: a thinly disguised rant

One of the significant changes in the area of Geocomputation over the past 20 years has been a shift in focus from top-down aggregate models to individual bottom-up approaches. This has been accompanied by an increased recognition of the role that the individual plays in driving key social processes that form a significant part of geographical systems (Batty, 2013, O'Sullivan et al, 2012). Whilst the acknowledgement that individuals are important components of these systems is not new in itself, the ability to chart the consequences of individual decisions and behaviours on geographical systems is. These new insights have been made possible through the development of new individual-based modelling methodologies enriched through the proliferation of micro-level population and economic data.

An individual-based method that has seen great uptake by researchers within Geocomputation over the past 20 years is agent-based modelling (ABM) (Macal, 2016). ABM advocates an understanding of social and spatial phenomena through simulation at the individual level. By creating heterogeneous individuals who can interact with other individuals and the environment, we can track the emergence of new patterns or trends across a variety of spatial and temporal scales. The emphasis within these models on the individual makes ABM a natural framework to apply within social and geographical systems as evidenced through the dazzling array of applications that are continually appearing, ranging from disaster relief (Crooks and Wise, 2013) to social epidemiology (El-Sayed et al, 2012). This popularity has been cemented by increases in computer processing power, data storage, developments in computer programming languages and easily accessible frameworks that enable rapid development of models with minimal programming experience.

While ABM offers a potentially powerful way both to simulate and to understand geographical systems, there remain several important challenges that researchers in ABM, and Geocomputation more broadly, need to address. Firstly, creating an agent-based model that can simulate the processes occurring in the real system requires the behaviours and actions of individuals, as well as environmental influences to be captured and represented. Current practice is lacking with the majority of 'behavioural' frameworks sharing more commonality with mathematics and econometrics than psychology. A more explicit link between ABM and behavioural frameworks is needed if we are to capture the complexity around decision-making and chart their consequences. Secondly, capturing this level of complexity requires a vast amount of individual-level data covering 'softer' factors such as feelings and opinions, data that more traditional quantitative research (and spatial science) has ignored. While the appearance of big data has opened up new avenues of research allowing highly complex models to be constructed that are enriched by new insights and understanding, how we extract value and make sense of these new forms of data presents a considerable challenge.

A final, and possibly the biggest challenge that ABM faces is that of calibration and validation. Creating realistic individual-based models requires a significant amount of data with a corresponding amount required to confidently calibrate and validate. As Heppenstall et al. (2016) note, there is some irony that by pursuing the disaggregation of data to the individual it becomes near impossible (at present) to rigorously calibrate and validate such models. However, even if the data were available, appropriate methods have not yet been established nor developed for measuring and analysing individual agents that are part of a large dynamic and non-linear system (Batty and Torrens, 2005; Torrens, 2010). This absence of robust calibration and validation measures has precipitated the criticism of ABMs as 'toy models'. Until researchers can fully evaluate these models against real world systems, it is unlikely that they will make the transition from academia into policy-making.

What is clear is that researchers now have the data and tools at their disposal to examine geographical systems in unprecedented individual-level detail thus creating new knowledge and understanding about how these systems evolved and what the consequences of future individual behaviours are

likely to be. The challenge for geocomputation is twofold: how can we use new forms of data to understand how real people shape and are shaped by geographical processes; and how can we realistically simulate these processes within our models?

Alison Heppenstall
School of Geography, University of Leeds

Breaking-out of the ivory tower

Over the past five years, a growth in geocomputational research has taken place away from academia, with many innovative new developments driven primarily by the commercial sector. In part this is their response to the opportunities arising through the emergence of big (geo)data in industry. These new forms of data challenge much of the pre-existing storage and processing infrastructure established at a time where contemporary "big data" did not exist. Unlike the traditional tasks of a database where a schema would be pre-defined and known, many applications exploring complex data sets require more flexible and adaptive technologies, and platforms such as Hadoop have been optimised for these purposes. There has been additional innovation from disciplines such as computer science around methods that use parallel optimisation, artificial intelligence, and supervised or unsupervised learning to translate data into useful insight. These methods may present a new epistemological approach within social science research (Kitchin, 2014) that challenges the frameworks of classical statistical inference long established.

Academia has been slow to keep pace and has not developed mechanisms that provide effective bidirectional dissemination of expertise and knowledge with industrial partners. This is regrettable because the potential benefits are not negligible. Beyond the pragmatic needs for innovation within the contemporary data economy, academia should be trying to engage more intensively with research activities of industry; conversely, industry should not underestimate the advantages of partnering with universities. Within the UK, the ESRC funded Consumer Data Research Centre (www.cdrc.ac.uk) makes an important step towards opening-up commercial data to academic users through secure data access facilities.

There has been significant growth in data science employment in roles requiring students who are geographically trained. For academia, this provides a significant constraint in attracting the most talented researchers and teachers (Rey, 2009). Although a challenge, the academic sector needs to do more to sell the benefits of research roles that include greater autonomy, more control over the destination and ownership of the outputs (including code), and the opportunity to work collaboratively across institutions without the shackles of protecting commercial interest. We take the view that academia needs to assume a more serious role as an incubator for innovation, where the knowledge, products and expertise developed as part of research activities can better be captured and have their exploitation supported in a way that generates a financial benefit to the researcher or teams involved. At the same time, academic institutions need to think carefully about how intellectual property generated by staff is captured and how these benefits may be shared, as well as how potential negative effects such as a reduction of open source development or reduced collaboration could be mitigated.

An increased interaction between industry and academia would make the latter more relevant to the former, and the former more useful and accessible to the latter, both to their mutual gain. We argue that the academic geocomputation community needs to engage more fully with some of the most

recent developments in the nascent field of Data Science. As others have argued elsewhere (Johnson, 2014), this conversation can be strengthened through training and education. A more targeted delivery of core geocomputation concepts and methods in the context of the Data Science world would demonstrate the value of incorporating space and geographical context into cases where geography is relevant to the (data) question at hand. A close inspection of some of the main textbooks (Schutt and O'Neil, 2013; Peng and Matsui, 2015; Pierson et al., 2015; EMC, 2015) and courses (Franklin, 2014; Irizarry and Hicks, 2016, John Hopkins University, 2016) on Data Science reveals there is a growing body of elements that remain remarkably consistent across all of them. This "basic curriculum" of Data Science broadly is composed of the following three areas: computational tools/software engineering, statistical methods, supervised and unsupervised machine learning, and data visualization. In all of these, there is little to no mention of explicitly spatial methods or wider considerations concerning their applications. At best, what we find are some examples of elementary mapping.

To address this deficiency, we propose a curriculum of what we term Geographic Data Science (GDS). The main elements that we believe could extend Data Science into an explicitly spatial domain are the following: spatial databases and file formats (e.g. GeoJSON, PostGIS); Exploratory Spatial Data Analysis (ESDA), in particular local measures; geodemographic analysis and regionalization techniques; spatial econometrics and geographically weighted regression; point pattern analysis; and cartography. These are not typical of a standard undergraduate method course in the social sciences yet they represent the sorts of techniques that need to be learned if future academics are to have the skillsets that are needed to engage with geocomputation within and beyond our ivory towers.

Alex Singleton and Daniel Arribas-Bel
Department of Geography and Planning, University of Liverpool

Geocomputation: conclusions, in way of catching breath

Looking over the abstracts from the first GeoComputation conference, two things leap out: the ahead-of-the-curve methodologies (machine learning; networks; web GIS; ABM; data-mining) and the breadth of applications areas. Geocomputation has been somewhat the victim of its own foresight in both: there are now tens of conferences in these methods and computational application areas. Nevertheless, one joy of the series is still being exposed to that breadth of techniques, both new and from other application areas.

Moreover, as Gahegan notes, the *idea* of geocomputation has proven even more important. Globally, staff, courses, and institutions are labelled geocomputational, or feel part of the subject. The *raison d'etre* of the series was to create a space for computation when quantitative geography was struggling against the "cultural turn" in geography. In many ways, its most important legacy is to allow people to hold their heads up and say "look, others elsewhere do this stuff; we should invest".

Nevertheless, in a world that has finally caught up, and where analysis and visualisation of spatial data are everywhere, it behoves us to ask "what now" for geocomputation?

First, there are issues to address. Our community gender balance is still poor and the traditional Anglo-American-Antipodean focus of the conferences is looking increasingly outdated. On teaching, Singleton and Arribas-Bel highlight the opportunity for clarifying geocomputation's unique-selling-points; we equally need to aim earlier, convincing children that coding is about more than making millions from an app and can be used to aid society. Finally, we need to manage our burgeoning

knowledge (>1,350,000 academic papers p.a.; Björk *et al.* 2009; many useful to geocompers). Brunsdon highlights Open Source data and techniques, and we should consider knowledge management to avoid re-creation and to identify which new and old techniques are useful, as well as their pitfalls.

More positive are our potential contributions to ongoing efforts in core areas. Industry, Singleton and Arribas-Bel note, is now investing in geocomputation far more than academia but we can still bring three things to the table. Firstly, *rigor*: we understand how analysis works in ways easily forgotten outside academia; those three spatial data daemons – the Modifiable Areal Unit Problem, Spatial Autocorrelation, and the Ecological Fallacy – still catch-out the naïve, as, in modelling, do Equifinality and Error Propagation. Secondly, *sympathy*: current solutions are driven by those with a narrow understanding of the world. Geocomputationalists are uniquely trained in the technical skills needed, but also a nuanced understanding of global systems. Thirdly, our breadth brings *imagination*: free from traditional subject boundaries, we can make unusual links and identify interesting opportunities.

Finally, we need to detail the future, as 21 years ago, and get at it, considering where spatially sensitive computing can make the world a safer, sustainable, and more satisfying place. Questions surround *data understanding and use*: Comber highlights re-negotiation of significance in a Big Data world, while Malleson notes the potential for dynamic data (and we might note for global social modelling); both demand thought on the social, political, and analytical *uses* of data. Beck, in 1987, appositely noted the important question is not how we predict the future using present parameters, but how we pick those needed to make it a *better* one. We also need to think more about how we track and display error and uncertainty associated with dynamic systems. As Heppenstall requests, *human experience* needs centring in our work: advances are waiting in capturing the emotional and belief-centred relationships between society and space. We also need to help develop a new politics of public duty and support in a world led increasingly by individual-level data and algorithms. As Longley notes, *space and place* are still key, but need updating with work on shared virtual and augmented realities, and their crossovers with the internet of things and telepresences. There's work needed on the emergent features of interconnected human systems – parallel economies and the influence of new and old media most urgently – but there's deep potential in understanding, visualising, and embedding the human experience as a node in a complex of interconnected flows. In AI, interactions with bots in complex social spaces, online and off, need elucidating, and geocomputation has a role to play in moving from machine learning to reasoning, as we attach structures and metaphors about the world to recognised objects. Finally, we have a place in *sustainability*: from resource optimisation to modelling planetary evolution and terraforming. In each area: human dynamics; experiences; uses of space; and interactions with the environment, we need those core principles: rigor, sympathy, and imagination, which promise insight and innovation in an exciting world of opportunities. If the last 21 years has seen the world catching up with us, the next 21 years should, with a fair wind and a strong heart, see us carry the world onwards.

Andy Evans
International Geocomputation Conference Series Steering Group
University of Leeds

References

Akhter S, Aida K, Chemin Y, 2010, "GRASS GIS on high performance computing with MPI, OpenMP and Ninf-G programming framework" *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science* **38**, 580–585

Anselin L, 1995, "Local indicators of spatial association – LISA" *Geographical Analysis* **27** 93–115

Antenucci J, 1989, Technical updates of geographic information. In *National Association of Counties Conference Workshop* (Cincinnati, OH)

Armstrong MP, 1995, "Is there a role for high performance computing in GIS?" *Journal of the Urban and Regional Information Systems Association* **7** 7–10

Asanovic K, Bodik R, Catanzaro BC, Gebis JJ, Husbands P, Keutzer K, Patterson DA, Plishker WL, Shalf J W, Williams S, Yelick KA, 2006, "The landscape of parallel computing research: A view from Berkeley" Technical Report UCB/EECS-183 (Electrical Engineering and Computer Sciences Department, University of California, Berkeley)

Barni M, Perez-Gonzalez F, Comesaña P, Bartoli G, 2007, "Putting Reproducible Signal Processing into Practice: A Case Study in Watermarking." 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (Honolulu, HI) doi: 10.1109/ICASSP.2007.367306

Batty M, 2013, *The New Science of Cities* (Cambridge MA, MIT Press)

Batty M, Manley E, Milton R, Reades J, 2013, "Smart London", in Bell S and Paskins J (eds.) *Imagining the Future City: London 2062* (London, Ubiquity Press) 31–40

Batty M, Torrens, PM, 2005, "Modelling and Prediction in a Complex World" *Futures* **37** 745–766

Beck MB, 1987, "Water quality modeling: A review of the analysis of uncertainty" *Water Re-source Research* **23(8)** 1393–1442

Björk B-C, Roos A, Lauri M, 2009 "Scientific journal publishing: yearly volume and open access availability", *Information Research* **14(1)** Paper 391 http://InformationR.net/ir/14-1/paper391.html [Accessed April 5 2017]

Black D, 1984, *Investigation of the possible increased incidence of cancer in West Cumbria. Report of the Independent Advisory Group* (London, HMSO)

Bond R, Kanaan A, 2015, "MassDOT Real Time Traffic Management System", in Geertman S, Ferreira J, Goodspeed R, Stillwell J (eds.), *Planning Support Systems and Smart Cities* (Cham, Switzerland, Springer) 471–488

Brunsdon C, 2016, "Quantitative methods II Issues of inference in quantitative human geography" *Progress in Human Geography* doi 10.1177/0309132516648020

Brunsdon C, Fotheringham AS, Charlton M, 1996, "Geographically weighted regression: a method for exploring non-stationarity" *Geographical Analysis* **28** 281–98

Brunsdon C, Singleton A, (ed.) 2014, *Geocomputation: A Practical Primer* (Sage, London)

Brunsdon C, Singleton A, 2015, "Reproducible Research: Concepts, Techniques and Issues," in Brunson C, Singleton A (eds.) *Geocomputation: A Practical Primer* (London, Sage) 254–64

Buckheit JB, Donoho DL, 1995, "WaveLab and Reproducible Research" Technical Report 474 (Department of Statistics, Stanford University)

Casetti E, 1972, "Generating models by the expansion method: applications to geographic research" *Geographical Analysis* **4** 81–91

Claerbout J, 1992, "Electronic Documents Give Reproducible Research a New Meaning" Society of Exploration Geophysicists Conference, New Orleans http://dx.doi.org/10.1190/1.1822162 [Accessed March 24 2017]

Comber A, Brunsdon CF, Charlton M, Harris P, 2016, "Geographically weighted correspondence matrices for local change analyses and error reporting: mapping the spatial distribution of errors and change" *Remote Sensing Letters*, DOI: dx.doi.org/10.1080/2150704X.2016.1258126

Couclelis H, 1998, "Geocomputation and space" *Environment and Planning B* **25** 41–47

Croitoru, A, Crooks A, Radzikowski J, Stefanidis A, 2013, "Geosocial gauge: A system prototype for knowledge discovery from social media" *International Journal of Geographical Information Science* **27** 2483–2508

Crooks AT, Wise S, 2013, "GIS and Agent-Based models for Humanitarian Assistance" *Computers, Environment and Urban Systems* **41** 100–111

Diao M, Zhu Y, Ferreira J, Ratti C, 2016, "Inferring individual daily activities from mobile phone traces: A Boston example" *Environment and Planning B: Planning and Design* **43** 920–940

El-Sayed AM, Scarborough P, Seemann L, Galea S, 2012, "Social network analysis and agent-based modelling in social epidemiology" *Epidemiologic Perspectives and Innovations* **9** 1

EMC, Education Services, 2015, *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data* (New York, Wiley)

Fischer MM, 2006, "Computational neural networks — tools for spatial data analysis", in Fischer MM (ed.) *Spatial Analysis and GeoComputation: Selected Essays* (Berlin, Springer) 79–102

Forer P, 1978, "A Place for Plastic Space?" *Progress in Human Geography* **2** 230–267

Fotheringham AS, Brunsdon C, Charlton M, 2002, Geographically Weighted Regression: the analysis of spatially varying relationships (Chichester, Wiley)

Franklin M, 2014, "Introduction to Data Science - UC Berkeley 2014" http://amplab.github.io/datascience-sp14/ [Accessed March 24 2017]

Gahegan M, 1999, "Guest Editorial: What is Geocomputation?" *Transactions in GIS* **3** 203–206

Gahegan M, 2000, "On the application of inductive machine learning tools to geographical analysis" *Geographical Analysis* **32** 113–139

Gentleman R, Temple Lang D, 2004, "Statistical Analyses and Reproducible Research" *Bioconductor Project Working Papers* Working Paper 2 http://biostats.bepress.com/bioconductor/paper2 [Accessed March 24 2017]

Gollini I, Lu B, Charlton M, Brunsdon C, Harris P, 2015, "GWmodel: An R Package for exploring spatial heterogeneity using geographically weighted models" *Journal of Statistical Software* **63** 1–50

Goodchild MF, 1992, "Geographical information science" *International Journal of Geo-Information Systems* **6** 31–45

Gorr WL, Olligschlaeger AM, 1994, "Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market modelling" *Geographical Analysis* **26** 67–87

Gould P, 1970, "Is *statistix inferens* the geographical name for a wild goose?" *Economic geography* **46** 439–448

Healey R, Dowers S, Gittings B, Mineter MJ, (eds.) 1997, "*Parallel Processing Algorithms For GIS"* (CRC Press, London)

Heppenstall A, Malleson N, Crooks AT, 2016, "'Space, the Final Frontier': How Good are Agent-based Models at Simulating Individuals and Space in Cities?" *Systems* **4** 9

Irizarry R, Hicks S, 2016, "Introduction to Data Science: BIO 260 and CSCI E107 - Harvard University" http://datasciencelabs.github.io/2016/ [Accessed March 24 2017]

John Hopkins University, 2016, "Coursera Data Science Specialization" https://www.coursera.org/specializations/jhu-data-science [Accessed March 24 2017]

Johnston RJ, Gregory D, Pratt G, Watts M, (eds.) 2000, *The Dictionary of Human Geography* (4th edition) (Blackwell, Oxford)

Jones JP III, Casetti E, 1992, *Applications of the Expansion Method* (London, Routledge)

Kitchin R, 2013, "Big data and human geography: opportunities, challenges and risks" *Dialogues in Human Geography* **3** 262–267

Kitchin R, 2014, "The Real-Time City? Big Data and Smart Urbanism" *GeoJournal* **79** 1–14

Kitchin R, 2014, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (London, Sage)

Koenker R, 1996, "Reproducible Econometric Research" (Department of Econometrics, University of Illinois)

Longley P, 2000, "Geocomputation", in Johnston RJ, Gregory D, Pratt G, Watts M (eds.) *Dictionary of Human Geography* (Fourth Edition) (Oxford, Blackwell) 296–7.

Lu B, Harris P, Charlton M, Brunsdon C, 2014, "The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models" *Geo-spatial Information Science* **17** 85–101

Lum K, Isaac W, 2016, "To predict and serve?" *Significance* **13** 14–19

Lynch MP, 2016, *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data* (Liveright, New York)

Macal CM, 2016, "Everything You Need to Know About Agent-based Modelling and Simulation" *Journal of Simulation* **10** 144–156

Malerba D, Esposito F, Lanza A, Lisi FA, Appice A, 2003, "Empowering a GIS with inductive learning capabilities: the case of INGENS" *Computers, Environment and Urban Systems* **27** 265–281

Malleson N, Andresen MA, 2015, "The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns" *Cartography and Geographic Information Science* **42** 112–121

Mather P and Openshaw S, 1974, "Multivariate methods and geographical data" *Journal of the Royal Statistical Society Series D* **23** 283–308

Mayer-Schönberger V, Cukier K, 2013, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (London, John Murray)

Miller HJ, Jiawei H, (eds.) 2009, *Geographic data mining and knowledge discovery* (CRC Press, New York).

Needle D, 2015, *Big Data's Value Much Larger than Specific Business Questions*. http://data-informed.com/big-datas-value-much-larger-than-specific-business-questions [Accessed February 6 2017]

O'Sullivan D, Millington J, Perry G, Wainwright J, 2012, "Agent-Based Models – Because They're Worth It?" in Heppenstall AJ, Crooks AT, Batty M, See LM (eds.) *Agent-based Models of Geographical Systems* (New York, Springer)

Olsson G 1969 "Inference problems in locational analysis", in Cox K, Golledge RG (eds.) *Behavioral Problems in Geography*, Northwestern Studies in Geography **17** 14–34 (Department of Geography, Northwestern University, Evanston, IL)

Openshaw S, 1991, "A view on the GIS crisis in geography, or, using GIS to put Humpty-Dumpty back together again" Environment and Planning A **23** 621–628

Openshaw S, Abrahart RJ, 1996, "GeoComputation", presented at the *First international Conference on GeoComputation*: http://www.geocomputation.org/1996/abs072.htm [Accessed March 24 2017]

Openshaw S, Charlton M, Wymer C, Craft A, 1987, "A mark I geographical analysis machine for the automated analysis of point data sets" *International Journal of Geographical Information Systems* **1** 359–77

Peng R, Matsui E, 2015, *The Art of Data Science*. https://leanpub.com/artofdatascience [Accessed March 24 2017]

Pierson L, Swanstrom R, Anderson C, 2015, *Data Science for Dummies* (New York, Wiley)

Pijanowski BC, Brown DG, Shellito BA, Manik GA, 2002, "Using neural networks and GIS to forecast land use changes: a land transformation model" *Computers, Environment and Urban Systems* **26** 553–575

Pradhan B, 2013, "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS" *Computers & Geosciences* **51** 350–365

Rey SJ, 2009, "Show me the code: spatial analysis and open source" *Journal of Geographical Systems* **11** 191–207

Rogan J, Franklin J, Stow D, Miller J, Woodcock C, Roberts D, 2008, "Mapping land-cover modifications over large areas: A comparison of machine learning algorithms" Remote Sensing of Environment **112** 2272–2283

Rogerson P, 1999, “The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic” *Geographical Analysis* **31** 130–47

Satish P, 2015, “Efficient parallel and distributed algorithms for GIS polygon overlay processing” (PhD Dissertation, Georgia State University). http://scholarworks.gsu.edu/cs_diss/98 [accessed Jan 14th 2017]

Schmidhuber J, 2015, "Deep learning in neural networks: An overview" *Neural networks* **61** 85–117

Schutt R, O’Neil C, 2013, *Doing Data Science: Straight Talk from the Frontline* (Sebastopol, CA**,** O’Reilly)

Schuurman N, 2000, “Trouble in the heartland: GIS and its critics in the 1990s” *Progress in Human Geography* **24** 569–590

Shi X, Kindratenko V, Yang C, (eds.) 2013, *Modern Accelerating Technologies for GIScience* (Springer, New York)

Stojanovic N, Stojanovic D, 2013, “High-performance computing in GIS: Techniques and applications” *International Journal of Reasoning-based Intelligent Systems* **5** 42–48

Torrens PM, 2010, “Agent-based Modeling and the Spatial Sciences” *Geography Compass* **4** 428–448

Vandewalle P, Kovačević J, Vetterli M, 2009, “Reproducible Research in Signal Processing.” *IEEE Signal Processing Magazine* **26** 37–47

Wiley EO, McNyset KM, Peterson AT, Robins CR, Stewart AM, 2003, “Niche modeling perspective on geographic range predictions in the marine environment using a machine-learning algorithm” *Oceanography* **16** 120–127

Wu N, Silva EA, 2010, “Artificial intelligence solutions for urban land dynamics: a review” *CPL bibliography* **24** 246–265

Yann L-C, Bengio Y, Hinton J, 2015, "Deep learning" *Nature* **521** 436–444