Birkin, Geoinfor Geostat: An Overview 2013, 1:1
http://dx.doi.org/10.4172/2327-4581.1000e101

Geoinformatics & Geostatistics: An Overview

Editorial

A SCITECHNOL JOURNAL

Big Data Challenges for Geoinformatics

Mark Birkin[1*]

It has been said that scientific research is now entering a 'fourth paradigm' [1]. While the earlier paradigms are characterised by experimentation and reasoning, the latest approaches are strongly driven by the availability of data at an unprecedented scale, and by the computational resources with which to extract the maximum value. Perhaps the classic exemplar discipline is bioinformatics, in which the human genome project has allowed a complete representation of the micro-level components and interactions which make up our physical bodies. The remaining challenge for scientific research, which is of course much easier to state than achieve, is 'simply' to deploy the methods, resources and imagination to discover the meaning in these rich streams of raw data. Major journals like Nature, Science and PNAS are now publishing the results of such data-driven research in almost every issue it seems (see for example the genomics special issue of Nature, 470 or the Big Data News Special, Nature 455).

These trends are of great importance to geographers, who share an interest with the biological scientists in mapping and deconstructing patterns in the real world, albeit at a somewhat more macro scale. Concepts such as virtual sensing, open data, and crowd-sourcing are now becoming commonplace. Sensors on the streets of every neighbourhood in the US and Western Europe are able to monitor traffic flows, pollution, precipitation and temperature. Where data continues to be collected by public institutions and agencies, such as the Ordnance Survey in the UK, these organisations are increasingly encouraged to open their holdings for public use and commercial exploitation. Crowd-sourcing, or volunteered geographic information [2], refers to the ability to generate data from the bottom up, for example by inviting and capturing opinions via a website or mobile phone applications - an excellent illustration is provided by research at the University of London examining variations in mood across space and time (www.mappiness.org). Such new data sources are a powerful complement to conventional sources such as the national censuses and behavioural surveys which have provided the bedrock for so much research over the last fifty years. Indeed the same trends have been remarked in other social science disciplines, notably sociology [3].

The question which now arises is how to capitalise, with suitable haste, on the opportunities which these new sources provide. One of the difficulties here may be the need to stitch together data which are many and varied, rather than trying to mine intensively from a small number of sources. We still need to find models which articulate the behaviour of individuals, organisations and systems; perhaps to find new ways of calibrating and validating these models; to visualise the results of our endeavours, alongside other ways of sharing our analytical frameworks and the outcomes from particular experiments or enquiries. Whether these requirements are generally recognised is debatable, even when the appropriate methods already exist or are in advanced state of development. Take for example the case of agent-based models (ABM) which have progressed rapidly in recent years. ABM provide an excellent way to represent geographical systems and to build relationships from the bottom-up, and this approach fits well with new data-driven paradigms which could provide the means to generate vast populations of realistic agents and to mimic their behaviour with robust observations from the real world. And yet it remains the case to an alarming degree that the most gifted exponents of the new generation of modellers are content to experiment with idealised toy systems. Still it seems that for every conference paper or journal article that actually attempts to validate a model there are ten which are content to bemoan the difficulties in doing so, or to identify further development as a headline for future research. As a community, I believe we need to show much more ambition in adapting our research to the 'grand challenges' of social science, and in particular to develop both deeper understandings and predictive models of how social entities operate at all scales, from individuals and communities, institutions and organisations, to governments and nations.

Here, ABM can be construed as one of the major elements in a broader armoury of geosimulation techniques which seeks to provide new model-based representations of real social systems at a real world scale. Geosimulation seems perfectly equipped to rise to the challenges and opportunities of fourth paradigm research. In recent work at the University of Leeds we have sought to demonstrate this in a number of ways. For example, we built a model of a new commuting charge in one England's major conurbations (the Manchester Congestion Charge) and then crowd-sourced the response of local people to calibrate that model, with a view to further articulation and evaluation of policy options [4]. Working with a local regeneration agency, we composed models of local household migration patterns which were embedded not just in census data but richly informed by massive commercial market research data sets with lifestyles, behaviour and attitudes as well as longitudinal spatial intelligence [5]. In our most recent work we consider the potential value in social media data ('tweets') for understanding daily movement patterns and interactions within cities, which must be a fundamental prerequisite for effective simulations of the transmission of illness or disease, as well as an underpinning for models of housing, labour markets, crime, transport, retailing and education, to name but a few [6].

The barriers to the roll-out of fourth paradigm research across the discipline remain considerable. One of the problems is sharing, whether that sharing relates to the data which fuel the work, the applications which provide its substance, or the publications which report its outcomes. Despite our earlier remarks, it is still common for organisations to try and protect their data, and this still happens in academic research where the data from particular studies are not always published or made generally available. It is even less common for authors to publish their codes and algorithms, although the advent of open source software languages, in particular R, is starting

***Corresponding author:** Mark Birkin, Room G06, East Building School of Geography, University of Leeds, Leeds, LS2 9JT, UK, Tel: +44 (0) 113 34 36838; E-mail: m.h.birkin@leeds.ac.uk

Received: October 12, 2012 **Accepted:** October 17, 2012 **Published:** October 19, 2012

doi:http://dx.doi.org/10.4172/2327-4581.1000e101

to make this more popular. Issues with charging and open access to publications are well-documented. A second layer of difficulties relates to ethics. The social databases of the future will increasingly be reliant on personal data of one form or another, and at present the rules and conventions for dealing with data of this type are loosely defined at best. Attempts to extract value from tweets are a good example. Ultimately the raw materials here are private conversations and remarks, and as this content is not protected by law, researchers must make their own judgements about the levels of aggregation or censorship which are appropriate to maintain the privacy of individuals. Thirdly, there are other cultural and practical difficulties. In particular as I have hinted above, the challenge is 'grand', and as in other cases from the physical sciences it will increasingly require large research teams working together to make significant progress. This is one reason why sharing between groups is so important, but for individual researchers it surely implies that a new synthesis between ambition and academic rigour may need to be found. It cannot be sensible to continue the process of academic reductionism indefinitely, when the important challenges require a higher level integration of data, methods and problems.

It is my sincere hope that the new journal of Geoinformatics and Geostatistics will provide an outlet for new styles of fourth paradigm research and geosimulation. What we are talking about here requires much more than scaling up of known methods for effective applications. Calibration, validation and real world applications are important precisely because they will expose the limitations of existing approaches; throw up new evidence and fresh challenges which depend on imagination and invention for their solution. I also urge contributors to think very seriously about making available their codes, data, and ancillary publications alongside their peer-reviewed academic articles as we seek to promote the development our discipline at a faster rate than ever before.

Geosimulation research at Leeds is supported by the Geospatial Data Analysis and Simulation (TALISMAN) node of the ESRC National Centre for Research Methods. For more details, please visit www.geotalisman.org.

References

1. Bell G, Hey T, Szalay A (2009) Beyond the Data Deluge. Science 323: 1297–1298.
2. Goodchild M (2007) Citizens as Sensors: the World of Volunteered Geography. Geo Journal 211-221.
3. Savage M, Burrows R (2007) The Coming Crisis of Empirical Sociology. Sociology 41: 885-899.
4. Birkin M, Malleson N, Hudson-Smith A, Gray S, Milton R (2011) Calibration of a Spatial Interaction Model with Volunteered Geographical Information. Int J Geogr Inf Sci 25: 1221-1239.
5. Jordan R, Birkin M, Evans A (2012) Agent-based Modelling of Residential Mobility, Housing Choice and Regeneration: Agent-Based Models of Geographical Systems, Springer, 511-524.
6. Malleson N, Birkin M (2012) Estimating Individual Behaviour from Massive Social Data for An Urban Agent-Based Model, 8th Conference of the European Social Simulation Association, Salzburg.

Author Affiliation

Top

[1]School of Geography, University of Leeds, Leeds, LS2 9JT, UK