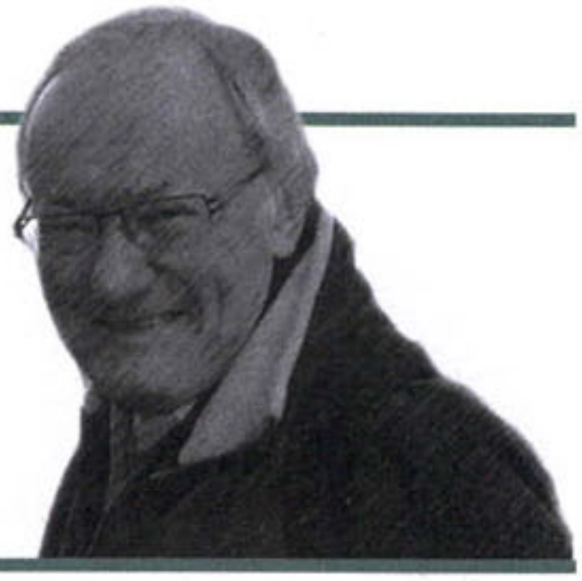
Big data; big issues



MICHAEL BATTY

is professor of planning at University College London and the author of *Cities and Complexity*

IN MY OPINION, geography is about to be transformed into a subject that deals with deluges of data – 'big data' as it's being called – and geographers need to begin to develop techniques to both analyse and assess the quality of such data, and to debate the issues surrounding its use.

Big data is largely generated automatically and routinely using machines that sense and sample activities in geographical space. Once the machinery has been set up to record and store such data digitally, the data streams only end when the machines are turned off.

In the past, most geographers' data were handcrafted, gathered using expensive, time-consuming surveys that we ourselves undertook. Now, however, many of our activities and actions are captured by sensors that we own, such as those in smart phones, or by computers and related devices embedded directly into both the built and natural environments. Quite suddenly, we're being overwhelmed by streams of data that can flow forever.

Big data isn't quite the same as 'big science', a term fashioned during the mid-20th century to encapsulate the organisations and equipment required to explore the sub-atomic world. The world of big data is much more like that of the human genome project, in the sense that it's very much part of the world in which we live, and the technical challenges relate to its storage, access and mining.

But these huge volumes of data are so recent that we have trouble even articulating the challenge. For example, in my group at University College London, we have six months of swipe-card data from Transport for London's Oyster Card scheme – something like one billion records. The system routinely captures some seven million or so trips each day using Tube, overground and bus systems. Seven million a day scales to about 50 million a week, to some 200 million a calendar month, which is nearly 2.5 billion a year.

Think of this data stream as never-ending from now on. In five years, we'll have 15 billion records – but what will they tell us about the world?

For years, human geographers following the work of Swedish geographer Torsten Hägerstrand have sought to understand time budgets and space-time trajectories, but progress has been slow, largely because of the difficulties involved in getting data. Now, however, we can trace millions of people whose location and times of travel are recorded digitally.

Of course, there are still powerful limits, but this kind of space-time data is now everywhere. Email traffic is spatial and geographical, as is our access to websites. Social media are generating countless flows of information at the individual level, and much of it's available for spatial analysis, notwithstanding the real problems in extracting meaning from short text messages in Twitter feeds.

But our ability to make sense of this enormous world of digital information and its impact on our geographies is still in its infancy. We're trying to glimpse this future, but to do so effectively we need new ways to unlock these enormous tranches of data. We need to link different data, adding value to such synthesis, and we need powerful new methods for data extraction, mining and visualisation, and methods to gauge the data's quality and representativeness.

The other issue raised by the emergence of this virtual world is the question of privacy. Most data generated in the public and quasi-public sectors are subject to strict controls, and it's difficult to identify individuals; however, it's never impossible.

In the private sector, the controls are less strict. There are questionable elements within Google's StreetView data, and Facebook's controls on the use of personal data. These are but two of the many large global media corporations that collect individual data and are enticed by their mining of it and its potential monetary value. But any digital data with unique identifiers can increasingly be linked to different data sets through the judicious use of correlation analysis.

Given this context, the landscape of big data is full of pitfalls and idiosyncrasies. Data on everything, everywhere, any time may be changing the world, but we need to figure out how good these data are and how ethical it is to use them, and to put them under the scrutiny of the various techniques we honed to tackle these issues in the bygone era of small data.

The advent of big data poses enormous challenges for geography, but also enormous opportunities. More and more, the art of mining and making sense of these data will be a place-related synthesis, which is what geographers are good at.

The need to acquire new tools to mine, visualise and communicate such data, and to extract meaningful geographies from them, provide an urgent agenda not only for research and practice, but also for education. If geography and geographers are to respond to this challenge, it's essential that they regenerate their curricula and interest in these new ways of articulating the world we've built. G