Dialogues in
Human Geography

Commentary

Dialogues in Human Geography
2014, Vol. 4(1) 45–49
© The Author(s) 2014
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2043820614525709
dhg.sagepub.com
SAGE

Spatial science – Looking outward

Chris Brunsdon
National University of Ireland Maynooth, Ireland

Abstract

When reviewing quantitative content in the geography curriculum, amongst other things it is important to review developments in data analysis outside of the discipline of geography. In this response to the paper by Johnston et al. (2014), a number of such developments are considered. In particular, the issues of big data, data journalism, reproducibility and statistical inference are discussed. In conclusion, it is argued that all of these would make some kind of positive contribution to the geography curriculum, providing in the words of Johnston et al. (2014) 'an important role in the formation of an informed citizenry in data-driven, evidence-based-policy societies'.

Keywords
big data, data journalism, geography curriculum, reproducibility

Like Johnston et al. (2014), I believe that the focus on the thinking underlying the quantitative revolution of the 1960s in modern texts on geographical thought and practice does, at least implicitly, lead to a misrepresentation of current quantitative geography – exemplified in Cresswell's (2013) statement that, 'The world of the spatial scientist is inhabited by a particular kind of imaginary person called a "rational being" . . . [such as] "rational economic man"' (p. 103). A realistic understanding of the role that 'spatial scientists' could play in current geographical debates can only be gained by acknowledging that this viewpoint does not typify much of their current thinking. Johnston et al. (2014) encapsulate current practice well by noting that 'much spatial science deploys place rather than space as its key geographical concept and there is rarely any mention of positivism and its basic tenets' (p. 6).

This is true both of exploratory data analysis – where emphasis is often placed on identifying places associated with unusual data patterns – and also with *local statistics* such as geographically weighted regression (Brunsdon et al., 1996) or *local Moran's I* (Anselin, 1995). However, an understanding of the changing position of spatial science does not depend solely on noting the change in approach of spatial scientists. Attention must also be paid to changes in thinking and technological developments in the world beyond spatial science and its practice within geography as a discipline. It is certainly the case that the 'outside world' has changed in terms of the way that agencies collect geographical data, and in terms of peoples' expectation of how this data may be used. In addition, ideas and debates in cognate disciplines, such as statistics, have also changed notably since the days of the

Corresponding author:
Chris Brunsdon, National University of Ireland Maynooth, Maynooth, Ireland.
Email: christopher.brunsdon@nuim.ie

original quantitative revolution. Current day spatial scientists are aware of these changes, and these have also influenced current thought and practice.

Whilst Wyly (2014) observes that the paper by Johnston et al. (2014) is written ‘by geographical scholars for an audience of scholarly geographers’, it is quite clear that the curricula of UK geography degree programs will affect the lives of many who, having gained degrees in geography, may well have to apply their knowledge in fields ranging beyond scholarly activities. For this reason, I would like to focus on these ‘external’ changes, and consider their implications for the geographical curricular in the United Kingdom. I will consider these under a number of headings – some external to academia and some within other areas of academia.

Debates external to academia

Big data

It is difficult to consider quantitative or ‘evidence-based’ approaches at the time of writing without reference to big data – and the situations in which the concept of big data is creating quantitative geographies – a quantitative spatial revolution outside of academic geography. As noted by Wyly (2014), this implies a sea change not just in the *amount* of data – some quantity of data has always been ‘big’ in the sense that standard statistical software had difficulty handling it – but in the way data are collected, in the people who collect the data, in the people who supply the data and in the motivation for the data collection.

This leads to the cautionary note in the discussion by Wyly – which I cannot really add to – but also to a number of other issues. If one is using big data to investigate behaviour a notable shift is away from the *designed experiment* (Stigler, 1992), where a carefully selected set of cohorts are studied in order to discover the effect of one particular factor on some aspect of behaviour. Although big data sets are large, one generally has to take what is volunteered – which may not match the careful sampling procedure above. There are both quantitative and qualitative issues here – from a quantitative viewpoint, highly imbalanced designs are frequently ineffective at detecting differences between different groups in the sample – and of course, biased samples can give modelling results that are simply wrong – see, for example, Brunsdon and Comber (2012). However, from a qualitative viewpoint, this suggests that there is a need to understand the sampling process in a broader sense – including an understanding of the actions of those collecting the data and those (possibly unconsciously) supplying it. Letting go of the ‘designed data collection’ paradigm can only be done with the effort of attempting to understand the new process of data collection.

Related to this, a large group of generators and users of big data exist in the commercial sector – and in many cases the data are treated as confidential to a specific organisation or are only available at a very high price. For academic research, this sits uneasily in an environment where those managing universities are responding to calls to ‘manage their finance effectively’ (Universities UK, 2013). Furthermore, the problem is then compounded in the United Kingdom by the threat to the continuation of the UK census, which as well as being an important resource for many quantitative studies in geography, is also a cost-effective one.

As well as the methodological and cost issues of big data, perhaps the greatest challenge (or possibly threat) comes from those claiming that big data is a substitute for theory in social science (e.g. Mayer-Schonberger and Cukier, 2013). In fact big data requires new theory, rather than an abandonment of theory – and that even if one were to accept the extreme assertion that society should ‘shed some of its obsession for causality in exchange for simple correlations’ (Mayer-Schonberger and Cukier, 2013: 7), one cannot do this without realising that any observed ‘simple correlations’ depend heavily on the data collection process. Perhaps in their enthusiasm for big data, the proponents of new forms of empiricism have forgotten to consider Simpson’s paradox – where patterns that appear in different subsets of data disappear when these subsets are merged, and a different pattern appears in the resultant data set (see e.g. Wagner, 1982). This has been around for some time and isn’t conditional on the size of the data set!

Data journalism

The term 'data journalism' is a relatively new one but is another area in which the visualisation and analysis of spatial information has become the focus of attention. In the United Kingdom, newspapers such as The Guardian offer 'crash courses' in the topic.[1] Most examples of data journalism rely on graphical approaches and, so-called *infographics*, some examples of which come under scrutiny from the statistical data visualisation community: See Gelman and Unwin (2013) and its responses for a lively discussion. Although not necessarily denying the validity of these approaches Gelman and Unwin argue that the statistical visualisation and the infographic have quite different goals. Quoting from their paper:

> On the statistical side, data analysts and statisticians are interested in finding effective and precise ways of representing data, whether raw data, statistics or model analyses ... On the Infovis side, computer scientists and designers are interested in grabbing the readers' attention and telling them a story. (p. 3)

However, the authors do identify a common purpose:

> The most general goals we can think of in data display are *discovery* [linking to the statistical goal] ... and *communication* [linking to the information visualization goal] ... These can go together—we want to communicate our discoveries! (p. 9)

Perhaps spatial science should combine these goals for spatial information – this is an area that should be explored, and the underlying principles be outlined in a quantitative geography syllabus.

Although big data has been identified by many as a key issue for latter day spatial scientists, it could be argued that the engine behind data journalism is not so much big data as *open data*, and the combination of several sources of data (mashups). Here, geography is often key – diverse data sets are often combined via spatial referencing – for example, linking levels of deprivation with rates of certain types of crime on an area-by-area basis, and a quantitative component in the geography curriculum allows participation in a dialogue relating to such analyses. How many data journalists are familiar with the *ecological fallacy* (Greenland and Robbins, 1994 or the *Modifiable Areal Unit Problem* (Gehlke and Biehl, 1934; Openshaw, 1983)?

As well as some presenting potential for dialogues, there is some technical common ground between data journalism and data science. For example, in the *Data Journalism Handbook* (Gray et al., 2012), the statistical programming language 'R' is advocated by a number of data journalism practitioners – indeed it is described as a 'Swiss army knife' of data visualisation and analysis. R is also taught as part of the degree course in a number of UK universities (e.g. Bristol, Leicester and Liverpool as part of either bachelor's or master's degrees). In addition, other software for handling and visualising geographical information (QGIS and ArcGIS) is also mentioned and again appears in the geography degree program curricula of many UK universities. Although these universities are perhaps in the vanguard of innovators, such subject matter could be an important component of any undergraduate geography syllabus. Knowledge of such software, in conjunction with other geographical knowledge, could potentially contribute to a significant cohort of new data scientists and data journalists who were critically aware not only of the technical procedures but also of the underlying geographical issues and debates.

Reproducibility

An issue gaining importance within statistical computing is that of *reproducible research* (Claerbout, 1992; Knuth, 1984) – essentially the idea that the ultimate product of academic research is the paper along with the full computational environment used to produce the results in it, such as the code, data and so on. The aim is that together these can be used to reproduce the results, scrutinise the arguments made on the basis of the data analysis and create new work based on the research. Although the term originated in the field of geophysics, it has subsequently been widely adopted in a number of disciplines – and certainly has relevance in the area of climate data analysis – where recent contentions such as *Climategate* called into question the results of statistical analysis appearing in some publications. In a report on this issue, chaired by Sir Muir Russell (2010), although finding that the rigour and honesty of those involved was upheld, the report stated that:

> We believe that, at the point of publication, enough information should be available to reconstruct the process of analysis. This may be a full description of algorithms and/or software programs where appropriate. (p. 104)

In addition, this issue is identified as important in social science – for example, an early advocate is Gary King (King, 1995) – indeed, the author of the extended commentary to this paper (Wyly) makes his own data and code available[2] citing King's paper and this principle. For human geographers using quantitative approaches, this is an important idea – if policy decisions are to be made on the bases of quantitative data analyses, it is not unreasonable that the data and computations underlying such analyses be made visible.

However, this presents new methodological issues – if the data are to be made publicly accessible, this requires that it is made open and that it is practical to distribute. The second point is perhaps particularly relevant of working with big data – even if when such data are made publicly available, this is usually via an application programming interface, and may involve distributing a random sample of data (which may not be identical when different users access the data) or a moving temporal window. Anyone analysing geographical data may at some point be answerable for the published results of their analysis – and subsequent conclusions drawn – and for this reason an awareness of the points outlined above should also be a key learning objective for any course on quantitative modelling or analytical techniques in geography.

Issues in statistical inference

A final area that I would like to focus on is that of statistical inference. This may seem more academic, and perhaps less current than the other issues I raise, but it is nonetheless a key issue in spatial science. The role of statistical inference is to assess the degree to which collected data supports theoretical hypothesis or to make statements about the calibration of theoretical models. Any discipline claiming to be evidence based must surely have the richness to critique the mechanism linking data to evidence.

However, the subject of statistical inference is itself the subject of some soul-searching within the statistical community, made evident through the rise of Bayesian approaches and the questioning of common practice in significance testing (Nester, 1996; Salsburg, 1985). Others seek to analyse the behaviour of those carrying out the tests (Marewski and Olsson, 2009). These arguments can certainly be applied to spatial statistics, and hence are of concern to the related area of spatial science – and to students whose curricula address these topics.

Finally, as well as the inferential techniques, those considering inference in geographical quantitative models should also be aware of critiques of existing quantitative models, such as Wall's 2004 critique of conditional autoregressive and simultaneous autoregressive regression models (Wall, 2004) – common tools of 'old school' spatial science. As well as debates on the inferential tools used, there are debates on the underlying models – I do not see such debates as problematic, but evidence of a vibrant community of users of geographical data, and geographical models, developing a discipline through self-examination. However, much of this is happening without input from geographers.

Concluding comment

In summary, many dialogues have been taking place for quite some time – and maybe geography scholars have been conspicuous by their absence. I therefore welcome this call for geographers to 'appreciate the underlying principles of quantitative analyses and their important role in the formation of an informed citizenry in data-driven, evidence-based-policy societies' and look forward to contributions that such awareness brings to future debates.

Notes

1. Web address: http://www.theguardian.com/guardian-masterclasses/data-journalism-course.
2. Web address: http://www.geog.ubc.ca/∼ewyly/replication.html.

References

Anselin L (1995) Local indicators of spatial association – LISA. *Geographical Analysis* 27: 93–115.

Brunsdon C and Comber L (2012) Assessing the changing flowering date of the common lilac in North America: a random coefficient model approach. *GeoInformatica* 16: 1–16.

Brunsdon C, Fotheringham AS and Charlton M (1996) Geographically weighted regression: a method for exploring spatial non-stationarity. *Geographical Analysis* 28: 281–298.

Claerbout J (1992) Electronic documents give reproducible research a new meaning. In: *Proc. 62nd Ann. Int. Meeting of the Soc. of Exploration Geophysics*, pp. 601–604.

Cresswell T (2013) *Geographical Thought: A Critical Introduction*. Chichester, UK: Wiley/Blackwell.

Gehlke CE and Biehl K (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association* 29: 169–170.

Gelman A and Unwin A (2013) Infovis and statistical graphics: different goals, different looks. *Journal of Computational and Graphical Statistics* 22: 2–28.

Gray J, Chambers L and Bounegru L (2012) *The Data Journalism Handbook: How Journalists Can Use Data to Improve the News*. New York, NY: O'Reilly Media.

Greenland S and Robins J (1994) Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. *American Journal of Epidemiology* 139: 747–760.

Johnston R, Harris R, Jones J, et al. (2014) Mutual mis-understanding and avoidance, mis-represen tations, and disciplinary politics: spatial science and quantitative analysis in (UK) geographical curricula. *Dialogues in Human Geography* 4(1): 3–25.

King G (1995) Replication, replication. *PS: Political Science & Politics* 28: 444–452.

Knuth DE (1984) Literate programming. *The Computer Journal* 27: 97–111.

Marewski JN and Olsson H (2009) Beyond the null ritual: formal modeling of psychological processes. *Journal of Psychology* 217: 49–60.

Mayer-Schonberger V and Cukier K (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Boston, MA: Houghton Mifflin.

Nester MR (1996) An applied statistician's creed. *Applied Statistics* 45: 401–410.

Openshaw S (1983) *CATMOG 38: The Modifiable Areal Unit Problem*. Norwich, UK: Geo Books.

Russell M (2010) *The Independent Climate Change E-mails Review*. East Anglia: University of East Anglia.

Salsburg DS (1985) The religion of statistics as practiced in medical journals. *The American Statistician* 39: 220–223.

Stigler SM (1992) A historical view of statistical concepts in psychology and educational research. *American Journal of Education* 101: 60–70.

Universities UK (2013) *Universities UK submission to the 2013 Spending Round*, p. 25.

Wagner CH (1982) Simpson's paradox in real life. *The American Statistician* 36: 46–48.

Wall MM (2004) A close look at the spatial correlation structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference* 121: 311–324.

Wyly E (2014) The new quantitative revolution. *Dialogues in Human Geography* 4(1): 26–38.