Progress report

Progress in Human Geography
2016, Vol. 40(5) 687–696

DOI: 10.1177/0309132515599625
phg.sagepub.com
SAGE

Quantitative methods I: Reproducible research and quantitative geography

Chris Brunsdon
National Centre for Geocomputation, Maynooth University, Ireland

Abstract
Reproducible quantitative research is research that has been documented sufficiently rigorously that a third party can replicate any quantitative results that arise. It is argued here that such a goal is desirable for quantitative human geography, particularly as trends in this area suggest a turn towards the creation of algorithms and codes for simulation and the analysis of Big Data. A number of examples of good practice in this area are considered, spanning a time period from the late 1970s to the present day. Following this, practical aspects such as tools that enable research to be made reproducible are discussed, and some beneficial side effects of adopting the practice are identified. The paper concludes by considering some of the challenges faced by quantitative geographers aspiring to publish reproducible research.

Keywords
Big Data, computational paradigm, geocomputation, programming, reproducibility

I Reproducibility in research

A great deal of practical quantitative work in human geography relies on the analysis of data – and it is often the case that published results are the final exposition of a great deal of behind-the-scenes data collation, re-formatting, coding, statistical modelling and visualization. It might be said that although published articles in this area exist to outline underlying questions, and draw conclusions from the data analysis, the conclusions will depend greatly on the behind-the-scenes work as well. This is why those carrying out this work are generally listed as authors. However, although the publication itself is a platform for discourse and debate around its content, it is sometimes harder to incorporate the behind-the-scenes activities into such debate, despite the fact that it can also influence conclusions and recommendations.

The term *reproducible research* (Claerbout, 1992) is used to describe an approach which may be used to address this problem. Although not noted greatly by geographers at the time of writing (but see Brunsdon and Singleton, 2015), it has gained attention in a number of areas where quantitative data analysis is used, for example: statistics (Buckheit and Donoho, 1995; Gentleman and Temple Lang, 2004), econometrics (Koenker, 1996) and signal processing (Barni et al., 2007). It is argued here that there is a strong case for a focus on this topic in quantitative geography. The goal of reproducible research is that

Corresponding author:
Chris Brunsdon, National Centre for Geocomputation, NCG, Maynooth University, Iontas Building, NUI, Maynooth, Ireland.
Email: christopher.brunsdon@nuim.ie

complete details of any reported results and the computation used to obtain them results should be available, so that others following the same procedures and using the same data can obtain identical results. This article considers the relevance and implications of this for geographical data analysis and GIS. Although the idea was put forward over two decades ago, the need to adopt reproducible practices is more relevant than ever. It has been argued that in addition to the two 'classical' paradigms of science that were commonly acknowledged at the time of the Claerbout (1992) paper (Hey et al., 2009; Kitchin, 2014b), two further paradigms are emerging:

1. *Deductive* (mathematics and formal logic)
2. *Empirical* (data collection, statistical model calibration and testing of hypotheses)

In chronological order a third *computational* paradigm uses algorithmic approaches such as large-scale simulation (for example agent-based modelling; Heppenstall et al., 2012) as a tool to gain insight into complex systems. Next, a fourth *exploratory* paradigm is emerging (Kelling et al., 2009), typified by the use of 'data mining' or, more generally, data-intensive approaches to identify interesting (arguably useful?) patterns in very large and structurally complex data sets. This emergence is in part due to the fact that advanced data collection, measurement and observational technology have made it possible to collect very large data (but often 'messy' data sets), and parallel advances in computer technology, such as cloud computing, mean it is possible to process such data sets in efficient ways. As the two 'traditional' paradigms interact, there are interactions between all four of the paradigms listed. For example, large-scale simulations are a way of exploring the consequences of certain mathematical assumptions arising from deductive approaches.

One thing linking the newer paradigms is their reliance on computer code (either created by the researcher or a third party) as an enabling technology. In both of the newer paradigms, although important ideas may be articulated in published texts, distinct intellectual contributions are embedded in software *code* where the ideas are represented in their most detailed form. Given this, a full critical engagement with researchers working within these paradigms is inhibited if code is not available openly. This is generally the case for quantitative science and social science, and for digital humanities. Here attention will be focused on the implications for quantitative geography, geocomputation and geographical information science.

II Geographical examples of reproducible research

For geographers, a consideration of the implications of the computational and exploratory paradigms is key in making the case for reproducibility. In terms of the computational paradigm, there is already a long tradition of the use of this approach. Although pre-dating the time when the idea of a computational paradigm in science was more common currency, work such as Openshaw and Taylor (1979) exploring the variation in correlation coefficients as areal units change demonstrates its use impressively. A key idea in the paper is this exploration of variability, but a comprehensive and accurate record of how this was achieved lies in the underlying FORTRAN code. Further examples include those related to microsimulation (Clarke and Holm, 1987). Lovelace and Ballas (2013) modify microsimulation techniques to provide simulations guaranteed to produce integer-based weighting for iterative proportional fitting (Ballas et al., 2005), and again the key ideas are those reflected in code. In this case, a fully reproducible approach is taken – in a supplement to the main article by Lovelace and Ballas a document outlines the technique in detail, incorporating code written in R (R Core Team, 2015) used to implement the algorithm. This

enables others to interact with the algorithm specified, and either modify it or apply it in a different situation, but one sufficiently similar that the same analytical framework would be meaningful. A similarly open approach is found in Ren and Karimi (2012), who present a fuzzy logic approach to GPS-based wheelchair navigation – here a link is provided to Java code used to implement their proposed algorithm.

An epidemiological example may be found in Parker and Epstein (2011), which uses agent-based models to simulate disease transmission on a global scale. In discussion, the authors provide a detailed outline of the underlying code used and, in particular, consider and provide details to assist in reproducing the code (including several code chunks), again making it possible to understand the underlying model (the key idea embodied in the article) more thoroughly and consider the effects of relaxing or modifying the assumptions of the model by modifying the code and re-running.

Other articles, although not providing full reproducibility – as they do not make the *exact code used* available – do provide very detailed descriptions so that there is a strong chance that a third party could reproduce the results. Although arguably this implies that *full* reproducibility is not achieved, papers adopting this approach demonstrate some of the advantages outlined above. For instance, Bergmann (2013) combines quantitative and qualitative approaches to consider global geographies of carbon emissions from a number of perspectives. For the quantitative part, full details of input-output models are provided which could be used to reconstruct and run analyses. A very different paper by Wood et al. (2012) similarly provides highly detailed computational description – in this case of algorithms and data graphics with the appearance of being hand-drawn. Although the direct code to produce the results seen in the paper is not shared, an open source library of tools is made available. In terms of reproducibility of the algorithm discussed, the author of this article was able to recreate it in R, for example, producing the results shown in Figure 1.

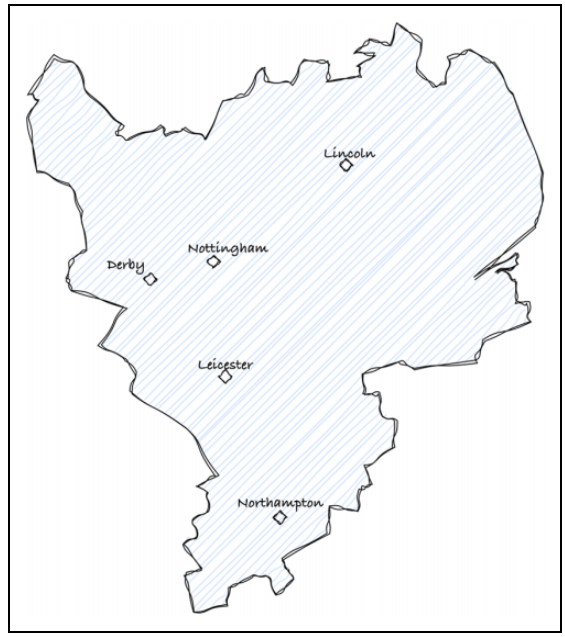

Figure 1. Map obtained by reproducing algorithm of Wood et al. (2012).

III A geographical case for reproducibility

Clearly, this idea is more practical in some areas of study than others, and resources are an important factor. It would not be feasible to re-run an entire census including data collection, collation and distribution, for example. However, in the area of quantitative human geography (assuming we accept census data 'as seen'!), and particularly spatial data analysis and GIS, it is a practical proposal in many cases.

The above may be seen as sufficient justification for reproducible research. However, if a more detailed case is to be made, the following scenarios (taken from Brunsdon and Singleton, 2015) help to reinforce the argument:

1. You have a data set that you would like to analyse using the same technique as described in a paper recently published by another researcher in your area. In that paper the technique is outlined in prose form, but no explicit algorithm is given. Although you have access to the data used in the paper, and have attempted to recreate the technique, you are unable to reproduce the results reported there.
2. You published a paper five years ago in which an analytical technique was applied to a data set. You now discover an alternative method of analysis, and wish to compare the results.
3. A particular form of analysis was reported in a paper; subsequently it was discovered that one software package offered an implementation of this method that contained errors. You wish to check whether this affects the findings in the paper.
4. A data set used in a reported analysis was subsequently found to contain rogue data, and has now been corrected. You wish to update the analysis with the newer version of the data.

Articles providing precise verbal description of algorithms are useful in these scenarios – as exemplified in the earlier examples – and it is certainly the case that this is a great improvement on vaguer descriptions that provide insufficient information to reproduce initial analyses. However, one could argue that the code itself is a much stronger aid to reproduction – a verbal description being prone both to incorrect interpretation and omission of necessary detail. In addition, there is the possibility that the code used in an article may contain an error, so that the precise description is in fact precise only in outlining what the author *thinks* it does – only the code itself will yield what it *actually* does. In most cases, the omission of such information is not done with malice aforethought on the part of researchers. Until the issue was raised in the article by Claerbout (1992) and those following, providing such detail was not considered standard practice in many disciplines. Indeed, some time later, few journals (none in geography, although this could be changing soon) insist that such precise details are provided, and it could perhaps be argued that there is some contributory negligence on their part.

Similarly, although it is usually required that researchers must cite the sources of secondary data, such citations often consist of acknowledgement of the agency that supplied this data, possibly with a link to a general website, rather than an explicit link (or links) to a file (or files) that contained the actual data used in the research, or details of any re-formatting of the data (including code) prior to analysis. However, both pieces of information allow published results to be critically assessed and scrutinized – ultimately leading to more trustworthy research conclusions.

IV The case for reproducible quantitative geography

The above is a general argument for reproducibility. However, one could ask whether this is relevant or practical for applications in quantitative human geography. In terms of relevance, it is worth noting that a great deal of analysis of social and economic data is inherently spatial – whether focusing on regional, local or street level – and that the results of such analyses are often used to inform policy-makers, and are used in decision-making processes. In many cases, the data being analysed is publicly available – for example, the US Census Bureau provide a number of APIs to access official statistics such as economic time series indicators and the decennial census for 1990, 2000 and 2010, the UK provides public access to census and reported crime data crime data, and Ireland provides access to Irish census data. However, not all reports or articles analysing this and

other publicly available data provide precise details of the analysis.

There are a number of arguments as to why such information *should* be provided. The first is a purely academic one – a useful and informed critical discourse of any analytical work can only take place when full details are provided. When the data analysis is a black box, it is difficult to either uphold or argue against any conclusions reached. One cannot tell whether the underlying models or techniques are appropriate or, even if they are, whether the underlying code or other computational approach faithfully reflects them. A second argument is one of accountability. Many quantitative studies inform policy decisions by governments and other institutions – different quantitative analyses with different outcomes could well lead to different policy decisions. Providing information not only about the sources of data used but also about the methods used to analyse the data is a key strategy of open government and democratic decision-making. As suggested earlier, this in turn leads to a more trustworthy approach – although this does not guarantee that an analysis is without error, it provides a mechanism where it is open to public scrutiny, so that the probability that any error is identified and corrected is notably increased. Also, relating to the earlier point, it implies that any assumptions made in the analysis are open to scrutiny, so that public discussion and debate regarding the basis of policy decisions is made possible.

A reminder of the relevance of this is provided through the recent controversy surrounding a paper by Reinhart and Rogoff (2010), whose published findings have been widely cited as an argument for fiscal austerity. However, in an article by Herndon, Ash and Pollin (2013), flaws were identified in the data analysis carried out in the paper. Quoting from the abstract of the latter article:

> We replicate ... and find that selective exclusion of available data, coding errors and inappropriate weighting of summary statistics lead to serious miscalculations that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies ... Our overall evidence refutes RR's claim that public debt/GDP ratios above 90% consistently reduce a country's GDP growth. (2013: 1)

This arose after a student, Thomas Herndon, unsuccessfully attempted to reproduce the analysis in Reinhart and Rogoff's paper as a coursework exercise. Investigations unearthed that the analysis was flawed – in part due to an error with an Excel spreadsheet. In this case measures were not taken to ensure reproducibility in the original paper – it took an amount of forensic computing to discover the problem. Following this, an errata was published (Reinhart and Rogoff, 2013), although Rogoff and Reinhart have defended their conclusions – if not their original analysis. However, the debate continues as authors of the critique continue to challenge a number of assumptions in the corrected analysis. Putting aside any criticisms I may have of the original paper, the outcome here is perhaps one of cautious optimism in that an open debate about the underlying analysis is now taking place – albeit after a great deal of public controversy. Again quoting from Herndon, Ash and Pollin:

> Beyond these strictly analytical considerations, we also believe that the debate generated by our critique of RR has produced some forward progress in the sphere of economic policy making. (2013: 279)

However, a reproducible approach here could have resulted in a smoother path to the final situation of public debate and a resolution of the erroneous analysis. Indeed, the spirit of the exercise set to the student was that of reproducing the published analysis.

V Achieving reproducibility

To address these problems, one approach proposed is that of *literate programming* (Knuth,

1984). This was initially proposed as a tool for documenting code, where a single file contained both the code documentation and the code itself. This was used to generate both a human readable document and computer readable content to generate software. The purpose of this was that the human readable output provided an explanation of the working of the program (and also neatly printed listings of the code), offering an accessible overview explanation of the program's function. However, such compendium files can also be used in a slightly different way, where rather than describing the code, the human readable output is an article containing some data analysis performed by the incorporated code. Tabulated results, graphs and maps are created by the embedded code. As before, two operations can be applied to the files – document creation, and code extraction. The embedded code is also visible in the original file. Thus information about both the reporting and the processing can be contained in a single document – and if this document is shared then a reproducible analysis (together with associated discussion) is achieved.

Examples of this approach are the NOWEB system (Ramsey, 1994), and the Sweave and Knitr packages (Leisch, 2002; Xie, 2013). The first of these incorporates code into LaTeX documents using two very simple extensions to the markup language. The latter two are extended implementations of this system using R as the language for the embedded code. Knitr also offers the possibility of embedding code into markdown – a simpler markup language than LaTeX – which facilitates very quick production of reproducible documents. The fact that R is used in the latter two approaches is encouraging for geographers, since R offers a number of packages for spatial analysis, geographical data manipulation of the kind provided by geographical information systems, and spatial statistics (Brunsdon and Comber, 2015). Furthermore, as R is open source software, the code used in any of these packages is also publicly available. Thus, not only is it possible to share high level data analysis operations, but also the code used to build the tools at the higher level.

Another possibility here is an approach using Pweave (Pastell, 2014) – a similar extension of NOWEB to embed Python code rather than R. Again, Python offers many tools for geographical data analysis, such as the PySAL package (Rey, 2015).

VI Beneficial side effects

Although much of the justification of a reproducible approach has been defensive, there are a number of benefits provided. Many of these occur as side effects when using the kinds of approach outlined above. In particular:

- Reproducible analyses can be compared: Different analytical approaches attempting to address the same hypothesis can be compared on the same data set, to assess the robustness of any conclusions drawn. In particular, a third party can take an existing reproducible document and add an alternative analysis to it.
- Methods are documented: One option with many reproducibility tools is to incorporate the code itself – as well as its outputs – in the documents produced. This allows for transparency in the way that results are obtained.
- Methods are portable: Since the code may be extracted from the documents, others may use it and apply it to other data sets, or modify it and combine it with other methods. This allows approaches to be assessed in terms of their generality, and encourages further dialog in terms of interpretation of existing data.
- Results may be updated: If updated versions of data used in an analysis are published (for example new census data), methods applied to the old data may be

re-applied and updated results compared to the original ones. Also, if the original data required amendment, an updated analysis could easily be carried out.

- Reports may have greater impact: Recent work has shown that papers in a number of fields, including reproducible analyses, have higher impact and visibility. This is discussed in Vandewalle, Kovačević and Vetterli (2009).

VII Challenges

The above sections argue that reproducible approaches offer a number of benefits. However, their adoption requires challenging changes in current practice. Perhaps one of the most notable is that the knitr, Sweave and Pweave approaches all require the use of *code* to carry out statistical analysis, visualization and data manipulation, rather than commonly adopted GUI-based tools, such as Excel. Unfortunately this is an inherent characteristic of reproducibility. After a series of point-and-click operations, results are cut and pasted into a Word document (or similar) and the link between the reported result and the analytical procedure is lost. It is perhaps no surprise that the Reinhart and Rogoff affair was seeded by an error in Excel.

Despite this, perhaps it is more realistic to consider ways in which the divide between GUI-based tools and reproducibility could be bridged than to propose such tools be abandoned. One possibility might be to provide GUI-based software in which every interactive event is echoed by a code equivalent, which is recorded. The recorded code could then be embedded in a document. One such tool that does this on a web-based interface is Radiant (Radiant News). However, it is perhaps also worth noting a general turn towards coding and away from GUI solutions in data analysis as indicated by the popularity of a number of books such as O'Neill and Schutt (2013) and McKinney (2012) – suggesting that there is a current wave of practitioners for whom the adoption of coding as a tool for data analysis does not imply a change of culture. Recent attendance at GIS conferences by the author would suggest, at least anecdotally, that these trends are reflected in geocomputation and geographical information science.

Other minor practical challenges also exist – for example, how can a sequence of random numbers in simulations be reproduced? However, many of these can be resolved by examples of 'best practice'. In the given example, random sequences may be made reproducible by noting that they are actually *pseudo-random* and specifying the code used to produce them, and the seed value(s).

However, a more significant challenge is created by the so-called 'Data Revolution' (Kitchin, 2014b) and the idea of *Big Data* – relating to the new paradigm of exploration and the search for empirical pattern, with implications of data mining and the search for patterns. Not only referring to the size of data sets, the term *Big Data* also refers to the diversity of applications, complexity of data and the fact that data is produced in a real-time 'firehose' environment where sensors and other data-gathering devices are streaming vast quantities of data every second. This is of importance to geographers applying quantitative techniques, since much of this data has a geographical component. The exploratory paradigm is not without controversy – while the computational paradigm could be viewed as working in co-operation with deductive and empirical approaches, some propose the exploration of Big Data as a superior competitor to theory-led approaches (see Mayer-Schonberger and Cukier, 2013, or Anderson, 2008), suggesting that working with near-universal data sets and identifying pattern supplants the need for theory and experiment. The title of the Anderson piece leaves little doubt as to the magnitude of the claim being made!

However, such boosterish claims have not gone unchallenged – notably, in the discipline

of geography, by Miller and Goodchild (2014), who argue, among other things, that there is still a need to understand the nature of the data being used and to discriminate between spurious and meaningful patterns. Kitchin (2014) warns of the risks of ignoring contextual knowledge in the analysis of Big Data. Although reproducibility in research involving Big Data analysis would not fully address any of these issues, it may be argued that it can provide a foothold. Giving precise details of assumptions in coding (for example, what kinds of patterns are being sought out by a particular data mining algorithm?) will certainly provide an entry point into dialogues addressing the issues raised above.

Despite this, currently many examples of reproducible research have used fairly 'traditional' approaches to data analysis, where a data set consists of a static file containing a rectangular table of cases by variables. More complex data poses less of a conceptual problem per se in terms of reproducibility – the challenge here is to devise appropriate analytical methods, but if that can be achieved then code can be created and reproducible research can be carried out in the ways outlined above. Similarly, diversity of applications presents no further conceptual difficulties for reproducibility. However, the real-time aspect does provide some challenges – clearly, even with the same code, two people accessing the same data stream at different points in time will not obtain identical results. One possibility might be to acknowledge that data used in a given publication is a static entity consisting of data obtained from a stream at a given point in time – and to time stamp and archive the data obtained and used in analysis at the moment it was carried out. Although it would be impossible for a third party to obtain identical data from the stream, and consequently impossible to obtain identical analytical results, it would at least be possible to see the code used to access the stream, note the time the stream was accessed, and access a copy of the data obtained at that time. This would also enable scrutiny of the representativeness of data – one contextual factor that may enable more meaningful analysis of Big Data.

VIII Conclusion

There are strong arguments for reproducibility in quantitative analysis of human geography data – not just for academics, but also for public agencies and private consultancies charged with analysing data that may influence policy. Achieving this in some situations is clearly within reach, although there are also some challenges ahead, as the diversity and volume of geographically referenced information increases. Arguably there is also a role for such methods in addressing the Big Data Revolution. However, the adoption of reproducible approaches does call for some changes in the practice of both researchers – in adopting reproducible research practices – and publishers – in providing a medium where reproducible documents may be easily submitted, handled and distributed.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 22 July 2015).

Ballas D, Clarke G, Dorling D, Eyre H, Thomas B and Rossiter D (2005) SimBritain: A spatial microsimulation approach to population dynamics. *Population, Space and Place* 11(1): 13–34.

Barni M, Perez-Gonzalez F, Comesaña P and Bartoli G (2007) Putting reproducible signal processing into practice: A case study in watermarking. *Proc. IEEE International Conference on Acoustics, Speech and*

Signal Processing. Available at: http://gpsc.uvigo.es/sites/default/files/publications/icassp07reproducible.pdf (accessed 22 July 2015).

Bergmann L (2013) Bound by chains of carbon: Ecological-economic geographies of globalization. *Annals of the Association of American Geographers* 103(6): 1348–70. DOI: 10.1080/00045608.2013.779547.

Brunsdon C and Comber A (2015) *An Introduction to R for Spatial Analysis and Mapping*. London: SAGE.

Brunsdon C and Singleton A (2015) Reproducible research: Concepts, techniques and issues. In: Brunsdon C and Singleton A (eds) *Geocomputation: A Practical Primer*. London: SAGE, 254–64.

Buckheit JB and Donoho DL (1995) *WaveLab and Reproducible Research*. Tech. Rep. 474, Dept. of Statistics, Stanford University.

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

Clarke M and Holm E (1987) Microsimulation methods in spatial analysis in planning. *Geografiska Annaler Series B, Human Geography* 69(2): 145–164.

Gentleman R and Temple Lang D (2004) Statistical analyses and reproducible research. *Bioconductor Project: Working Paper 2*.

Heppenstall A, Crooks A, See L and Batty M (2012) *Agent-Based Models of Geographical Systems*. New York: Springer.

Herndon T, Ash M and Pollin R (2013) Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics* 38: 257–279.

Hey T, Tansley S and Tolle H (2009) Jim Gray on eScience: A transformed scientific method. In: Hey T, Tansley S and Tolle K (eds) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research. Available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf (accessed 22 July 2015).

Kelling S, Hochachka WH, Fink D, Riedewald M, Caruana R, Ballard G and Hooker G (2009) Data-intensive science: A new paradigm for biodiversity studies. *BioScience* 59(7): 613–20. DOI: 10.1525/bio.2009.59.7.12.

Kitchin R (2014) Big Data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography* 3(3): 262–267.

Kitchin R (2014a) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1). DOI: 10.1177/2053951714528481.

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

Knuth D (1984) Literate programming. *Computer Journal* 27(2): 97–111.

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

Leisch F (2002) Dynamic generation of statistical reports using literate data analysis. In: Härdle W and Rönz B (eds) *Compstat 2002: Proceedings in Computational Statistics*. Heidelberg: Physika Verlag, 575–580.

Lovelace R and Ballas D (2013) 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems* 41: 1–11.

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

McKinney W (2012) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. New York: O'Reilly.

Miller HJ and Goodchild M (2014) Data-driven geography. *GeoJournal*. DOI: 10.10007/s10708-014-9602-6.

O'Neill C and Schutt R (2013) *Doing Data Science: Straight Talk from the Frontline*. New York: O'Reilly.

Openshaw S and Taylor PJ (1979) A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In: *Statistical Applications in the Spatial Sciences 21*. London: Pion, 127–144.

Parker J and Epstein J (2011) A distributed platform for global-scale agent-based models of disease transmission. *ACM Transactions on Modeling and Computer Simulation* 22(1). DOI: 10.1145/2043635.2043637.

Pastell M (2014) Pweave: Reports from data with Python. Available at: http://mpastell.com/pweave/docs.html (accessed 22 July 2015).

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org/ (accessed 22 July 2015).

Radiant News (2015) Introducing Radiant: A shiny interface for R. Available at: http://www.r-bloggers.com/

introducing-radiant-a-shiny-interface-for-r-2/ (accessed 22 July 2015).

Ramsey N (1994) Literate programming simplified. *IEEE Software* 11(5): 97–105.

Reinhart CM and Rogoff KS (2010) Growth in a time of debt. *American Economic Review: Papers and Proceedings* 100(May): 573–578.

Reinhart CM and Rogoff KS (2013) Errata: Growth in a time of debt. Harvard University, 5 May. Available at: http://www.carmenreinhart.com/user_uploads/data/36_data.pdf (accessed 22 July 2015).

Ren M and Karimi HA (2012) A fuzzy logic map matching for wheelchair navigation. *GPS Solutions* 16: 274–282. DOI: 10.1007/s10291-011-0229-5.

Rey S (2015) Python Spatial Analysis Library (PySAL): An update and illustration. In: Brunsdon C and Singleton S (eds) *Geocomputation: A Practical Primer*. London: SAGE, 233–254.

Vandewalle P, Kovačević J and Vetterli M (2009) Reproducible research in signal processing. *IEEE Signal Processing Magazine* 26(3): 37–47.

Wood J, Isenberg P, Isenberg T, Dykes J, Boukhelifa N and Slingsby A (2012) Sketchy rendering for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 18(12): 2749–2758.

Xie Y (2013) *Dynamic Documents with R and Knitr*. New York: Chapman and Hall CRC.