INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
VOL. 30, NO. 8, 1507–1521
http://dx.doi.org/10.1080/13658816.2015.1137579

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Establishing a framework for Open Geographic Information science

Alex David Singleton [a], Seth Spielman [b] and Chris Brunsdon [c]

[a]Department of Geography and Planning, University of Liverpool, Liverpool, UK; [b]Department of Geography, University of Colorado, Boulder, CO, USA; [c]National Centre for Geocomputation School of Environmental Sciences, National University of Ireland at Maynooth, Maynooth, Ireland

ABSTRACT

When conducting research within a framework of Geographic Information Science (GISc), the scientific validity of this work can be argued as highly dependent upon the extent to which the methods employed are reproducible, and that, in the strictest sense, can only be fully achieved by implementing transparent workflows that utilize both open source software and openly available data. After considering the scientific implications of non-reproducible methods, we provide a review of both open source Geographic Information Systems (GIS) and openly available data, before describing an integrated model for Open GISc. We conclude with a critical review of this embryonic paradigm, with directions for future development in supporting spatial data infrastructure.

ARTICLE HISTORY
Received 8 November 2014
Accepted 29 December 2015

KEYWORDS
Geographic Information Science; Geographic Information Systems; reproducibility; open source; Open Data

1. Introduction

There are multiple views of the peer review process in academic publishing. Peer review can be seen as a quality control mechanism, a means to enforce scientific integrity and ensure that publicly reported science is not conceptually or methodologically flawed. In short, peer review is a way to check the quality of published research. Peer review can also be seen as a gate-keeping mechanism. That is to make sure the language, format and presentation conform with disciplinary norms. As anyone who has worked with interdisciplinary teams will know, these standards and norms are often discipline-specific and have evolved around a basic tension in academic publishing. On the one hand, the social purpose of publishing is to ensure research enters the public domain in a form that allows replication by a third party (Fleck 1981, Jasny *et al*. 2011, Jasny 2013). On the other hand, research is hard work and the authors often feel a sense of commercial and intellectual ownership of outputs. While it has been argued that a guiding principle of Geographic Information Science (GISc) is an assurance of transparency in the assumptions and methods (Longley *et al*. 2010), in publishing, there can be a tension between ownership and transparency. This tension is navigated through standards that dictate articles in peer-reviewed journals have a sufficiently well-specified narrative on the sources and attributes of data, alongside those methods that would enable reproduction

CONTACT Alex David Singleton alex.singleton@liverpool.ac.uk

© 2016 The Author(s). Published by Taylor & Francis.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/Licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

of results. However, this is often accompanied by an implicit caveat that for full verification of any presented results to take place, a third party would also require access to both the data and software used to conduct such analysis. Within much of the geographic domain, science depends on systems, that is, results are generated through the application of software to data or via the development of algorithms (which in turn may depend on existing software/algorithms).

In this paper, we argue in favor of a shift in the norms around academic publishing of GISc for a model in which a journal-based 'publication' encompasses prose, code and where possible data as part of a 'workflow' that enable replication of analysis/presentation (Peng *et al*. 2006, Rey 2014). We argue that such a model evokes enhanced public accountability and reproducibility given that the workflows could be re-run and the output tested as part of both the review or dissemination process. Such a process would advance the discipline by reducing the cost of improving/implementing novel research. However, there are significant challenges to this idea: modern publishing infrastructure is 'text-centric' in that it emphasizes written convent over scientific reproducibility, and publishers/libraries often lack the resources to host archives of code and/or data. Licensing models also generally focus on software products or data and not scientific work; as such, new licensing models may be necessary to protect scholars' intellectual property. Finally, software can sometimes be closed source, and as such limit public dissemination; or data may also be either private or of a sensitive nature, and as such might prevent public disclosure. Others have argued that the current ownership structure and licensing arrangements surrounding academic published material provide an overarching constraint on openness and transparency (Tamber 2000); however, we do not focus on this specific issue in this paper. Instead, we present the case for a new model of open publishing in GISc oriented toward the social and scientific benefits of increased transparency and reproducibility in academic 'publishing' while respecting the tension between openness, reproducibility and intellectual property.

2. The dangers of non-reproducible science

The dangers of non-reproducibility were evidenced in 2009 when controversy arose surrounding the hacking of e-mail servers housing messages from a high-profile team of climate scientists at the University of East Anglia, UK (Campbell 2009). It was argued by critics that these e-mails provided evidence that undermined the research findings of the team involved, and that given their prominent position with government officials, cast doubt among skeptics about the overall robustness of the climate change mitigation policies that their work had informed. These issues were widely debated across the public press, with a key issue being that the researchers would not release their data, thus negating the possibility of replicating results (Campbell 2010). However, it transpired later that the meteorological data that was integrated into the models could not be repackaged for release because it had been supplied under a commercial license, thus denying redistribution to third parties. If these data had been supplied under an open license, the data and output of the models could have been made available and thus offered the potential to be checked by critics. A second example was illustrated recently in the field of economics and concerned a paper published by two eminent Harvard University Professors (Reinhart and Rogoff 2010). This presented findings that

gained wide international traction as key evidence for government austerity-based fiscal policies. However, 3 years after publication, and as part of a PhD student project at University of Massachusetts, Amherst, MA, USA (Herndon *et al*. 2013), a student found that the results could not be reproduced, and after requesting the original Excel spreadsheet from the authors, a significant coding error was spotted where a number of countries had accidentally been omitted from a key calculation. A final example is sourced from the field of cancer treatment where gene-based tests utilized for drug treatment selection were shown to be derived from erroneous research. Baggerly and Coombes (2009, 1310) note that 'poor documentation and irreproducibility can shift from an inconvenience to an active danger when it obscures not just methods but errors. This can lead to scenarios where well-meaning investigators argue in good faith for treating patients with apparently promising drugs that are in fact ineffective or even contraindicated'.

In different ways, the cited examples illustrate the importance of reproducibility, and specifically how access to input data sources can either help mitigate potentially erroneous conclusions being drawn from empirical observations or give the public greater assurance about the validity of presented work. Given these imperatives, improving methods of reproducible research is a topic gaining traction across a range of scientific disciplines (Stodden 2014), with examples arising, but not exclusive to bioinformatics (Gentleman and Temple Lang 2004, McMurdie and Holmes 2015), signal processing (Vandewalle *et al*. 2009), gene pattern analysis (Reich *et al*. 2006), acoustics (Kovacevic 2007), epidemiology (Peng *et al*. 2006) and economics (Baiocchi 2007). More generally, within both science and social science, there are a growing number of initiatives that aim to explicitly test the reproducibility of peer-reviewed research, for example, through the 'crowd sourced' Reproducibility Project (opensciencefframework.org) or the Reproducibility Initiative (www.scienceexchange.com/reproducibility).

Our paper is positioned within this literature on more general open reproducible science, describing a model that would be applicable for much GISc research. Furthermore, we argue that given a growing interest in the relevance of space to those practices of other disciplines (Warf and Arias 2009), including an expansion of Geographic Information Systems (GIS) activity outside of traditional enclaves (Goodchild and Janelle 2010), there is an urgency for the GISc community to revisit publication standards, and specifically, to reconsider how those data, methods, tools and analysis procedures that we utilize can be made more explicitly open and scrutable, thus engendering greater scientific transparency and reproducibility.

3. Open GIS

Contemporary GIS typically adopt either an open source or closed source model (Steiniger and Bocher 2009). Open source applications are available in source code format within the public domain, and under a number of licenses that typically permit different permutations of reuse, adaptation and redistribution for commercial or non-commercial purposes. These softwares are prevalently available at no cost, and development work is often completed by a community of programmers (von Krogh *et al*. 2003), although some have debated the extent to which open source software are group efforts (Krishnamurthy 2002). This model contrasts with closed software, where the

source code used to create the applications is not typically available in the public domain. However, some commercial closed source software vendors also release parts of their source code, or extensions (such as scripts) into the public under open licenses, and as such, represent a hybrid model.[1]

The number of open source GIS projects are expanding (Steiniger and Bocher 2009) aligned with a growing community of users supported by organizations such as OSGeo and the International Cartographic Association (ICA). One such example includes the joint collaboration network setup by these two organizations called 'Geoforall',[2] which represents at the time of writing 91 ICA-OSGeo Labs worldwide. Such labs encourage the use of Free and Open Source Software for Geospatial (FOSS4G) analysis (O'Brien 2014), offering a variety of interaction opportunities including open source training (Cheshire 2014) which are disseminated to a wide audience given the lack of licensing restrictions. For those successful projects, the number of users downloading software are also increasing greatly. Understanding the drivers of these trends is important. First, much of the software development is happening outside of the academic sector where 'publishing' simply means making code available online, and typically through a repository-like github.com. The motivations for the open source software community are interesting to consider with this regard; Lakhani and Wolf (2005) drew a random sample of open source projects that were under active development and conducted a census of finalized products on SourceForge, which is a code repository for collaborative programming projects. Their sample contained of 684 contributors (30% response rate), providing data on 287 different open source projects. Lakhani and Wolf (2005) found that respondents were overwhelming male (97.5%) with a mean age of 30 (SD: 8 years, the youngest contributor was 14, the oldest was 56). The contributors top three motivations were diverse, some wanted to improve programming skills (41.3% selected as a top three motivation), others found programming intellectually stimulating (44.9%), but the most common motivation was needing the collectively generated code for a separate commercial or non-commercial project (58.7%). Bitzer *et al*. (2007) argue that one of the most important factors leading to the creation of an open source project is perceived need, where projects are started to address gaps or shortcomings in existing software systems. For example, Yang and Lai (2010) studied the motivations of 200 English language contributors to Wikipedia and found that internal self-concept motivation was the most important predictors of the frequency of individual contributions. Yang and Lai define internal self-concept motivation as a form of positive feedback that occurs when a behavior is consistent with an individual's ideals and provides positive external feedback. People contribute to collective knowledge bases when they find the activity consistent with their ideals and they receive some praise for doing so from others. Thus, what matters for open source developers is not recognition in the form of citations, it is membership of a community and acting in a way that is consistent with their ideals. While belonging to a community is nice, academic professionals success is assessed using different metrics. That said, even without a clear return to effort, open source GIS is expanding rapidly and there are benefits to a tighter collaboration between academics and open source developers (Rey 2009). Although there are no current figures available for the latest releases. Sutton (2010) notes that a recent version of the QGIS (www.qgis.org) was downloaded around 100,000 times in the 4 months after release; which compared to a previous

version over the same time period was an increase of around 60,000 downloads, thus representing a massive expansion in user base.

3.1. *Open GIS and advanced spatial analysis*

Open source GIS projects encapsulate a broad range of software types, ranging from specialist tools used to complete analytic functions (e.g. to build a cartogram), through to multi-purpose desktop or web GIS that incorporate a range of functionality. An expansive review of the major Open GIS Desktop projects can be found in Steiniger and Bocher (2009) and Steiniger and Hunter (2012) so will not be repeated here. Instead, we focus on how advanced spatial analysis techniques are included in a wider class of open source GIS software that have developed as extensions to high level statistical and general purpose programming languages, or, as separate libraries that provide specific functionality accessible by a range of software languages. When a programming language is referred to as high level, this relates to the degree of abstraction from some of the more complex aspects of writing software; for example, how to handle memory and computer processing unit allocations; thus making the use of the language more accessible to a wider range of users. These classes of software are especially pertinent for open GISc and publishing. Since analysis is completed via a written language as opposed to mouse clicks, a workflow is preserved and easily replicated, the use of written languages makes an analysis to some minimal extent self-documenting.

The R (www.r-project.org/) software environment provides an example of this class of open source GIS. R is designed for statistics and data manipulation operations but has been extended through a range of additional spatial statistics and geovisualisation packages (Bivand *et al*. 2008). Packages are installed in addition to the basic installation of R to provide extended functionality. For example, the vast majority of spatial statistical techniques featured in traditional GIS can be implemented within R through an extension package such as 'spdep' (Bivand 2014). Additionally, the R graphing capabilities have been adapted to provide map-based outputs through packages such as 'maptools' (Bivand and Lewin-Koh 2013) and can even layer these outputs with other contextual data such as OpenStreetMap. R provides the ability to perform manipulation and statistical operations on datasets by writing commands in R syntax, and then running these through an interpreter that outputs the results from the instructions specified in the code. The ability to specify and report data manipulation, statistical and spatial analysis commands alongside outputted results creates more transparent and reproducible science; first because the code can be scrutinized for errors as part of the peer review process, and second, because any specific functionality employed by packages called by the code can also be checked at source code level for errors; given that these are also open source.

The same is also true of the Python programming language which has extensive spatial data handling and analysis libraries (e.g. PySal: geodacenter.asu.edu/projects/pysal; Pandas: pandas.pydata.org; GeoPandas: geopandas.org; Statmodels: statsmodels.sourceforge.net). In some sense, this use of text input (codes) rather than menu-driven commands of a graphic user interfaces (GUI) returns to the origins of GIS as a terminal/mainframe computer software. Arguably GUI have made access to the analytic functions of GIS more readily accessible to a wider audience of users as the desktop/GUI paradigm

has prevailed in the majority of software groups; however, we argue here that GUI in the context of social science limit reproducibility, as replicability of functions implemented is more difficult (although not impossible) to document, and furthermore, in GUI-based software, there is a tendency to hide the complexity of analytic functions from users. For example, in ESRI ArcGIS software, you can conduct a 'spatial join', aiming to create some combination of points, lines or polygons; however, the actual processes implemented for the different combinations of these spatial objects are reasonably hidden (although well documented) from the users. We would argue that this increases the potential for users to create results in error, as there is less of a requirement to think about the underlying processes that are being implemented.

3.2. *Balancing open and closed software*

At their most restrictive, the licenses that govern the use of open source software require that any product developed utilizing the associated code is also released under an equivalent open license. For example, under a restrictive licensing agreement, like the GNU Public License (GPL), anyone is free to use the code; however, any product that incorporates this code must also be fully open source. For example, the R software and packages are licensed under a range of licenses including (GNU, AL, BSD MIT), whereas Pysal is BSD. This potentially presents license compatibility problems when developers may wish to mix code from several existing products, perhaps with such variable licensing/distribution restrictions.

A further critical point is that the licenses are developed to govern the reuse of software, not services. In spatial analysis, the innovation is often in the application of existing methods in novel ways. For example, if a person develops a useful algorithm for processing data, they can restrict the use of the code in closed-source commercial software products, but they cannot restrict the commercial use of the code in service provision. This means that the creator of the method by publishing open source is essentially forfeiting the ability to capitalize on their innovation and/or is facilitating the commercialization use of their methods by others. It would be beneficial to open source GISc if licensing agreements protected intellectual property by restricting commercial use of methods. In other fields, such as Computer Science, patents are commonly used to protect the value inherent in intellectual property. In spite of this yet to be resolved issue around a fully open source publishing model, there are major issues associated with closed source practice.

Closed source software can lack transparency in the specification of analytic functions. For example, a GIS may implement the commonly used spatial interpolation technique of kernel density estimation, which can utilize a range of different kernel types or bandwidth selections. Each result would produce slightly different results, and without knowing which specific type of kernel or bandwidth that was used in an analysis, errors could be made when interpreting the output representations or these would be difficult to reproduce. Although some closed source GIS software vendors provide detailed specification of how algorithms are implemented within their tools, no suppliers provide information that would be considered equivalent to those insights that could hypothetically be garnered by examination of the source code directly.

This issue of transparency is also compounded by a risk that systematic errors might be introduced into analysis due to erroneous coding of specific functions. Without public domain source code access, software cannot be verified as error-free, and as such, a question over the validity of outputs may remain. Although not all closed source software tools are commercial, it might be argued that such errors would be unlikely in software that is sold given the potential financial implications of such mistakes, and that formal error checking processes are often built into the release schedules of most commercial software. Conversely, in open source GIS where source code is publicly accessible, functions could hypothetically be checked prior to implementation, thus mitigating potential errors in the interpretation of results. Furthermore, open source software is typically programmed by multiple developers, the frequency of which will differ between projects; applying the principle of 'many eyes' (Raymond 2001) where robustness in code aims to be assured by multiple users checking and fixing errors. However, the open source paradigm is not a panacea, and there is evidence to suggest that for many open source projects, development is actually completed by a more limited set of contributors, and as such, potentially undermines the extent to which code is checked for errors (Krishnamurthy 2002).

A final issue for reproducibility with closed source commercial GIS is that there is typically a cost associated with the use of software that restricts reproducibility to only those who can afford access. For some groups of users, such as many academics, this may be met with sector-wide licenses; however, these types of agreements are not necessarily applicable in all countries, or between different stakeholder groups. In particular, issues of cost can be acute for non-governmental organizations or not for profit groups where incomes may be restricted, yet may fall outside of any national agreements linked to central governments.

4. Openly available data

For analysis to be reproducible, third parties require access to data. This is cited as a challenge for reproducibility (Borgman 2012), with particular constraints noted in the area of subject confidentiality, conditions of use, restricting of access from commercial entities and time embargoes (Stodden 2010). For example, in the past, much UK social science data including the Census of the Population or administrative data have been restricted to certain groups of users such as academics or licensed third-party distributors. The 1991 Census of the Population for example was licensed centrally by the Office for Population Census Surveys and made available to academics for free at the point of use through restricted services and to other users through commercial resellers. For users outside of academia, or without the ability to pay for commercial licenses, this would have prevented access, and as such, it would have been very difficult to argue that any research based on these data were reproducible by all. We would however make the case that if data were not available within the public domain, this should not preclude a researcher adopting other aspects of an Open GISc framework.

However, although such constraints are still in existence with certain data as evidenced in the climate change example cited earlier (Campbell 2009), there is a more general movement toward much large public data being released with less restrictive licenses, which is driving a new paradigm in data sharing and reuse. Within a UK context, the term

'Open Data' has an explicit definition as relating to those data that have been released under the Open Government License for Public Sector Information (www.nationalarchives.gov.uk/doc/open-government-licence); however, this term is often also used more flexibly to refer to those data that are free from reuse restrictions or financial cost of acquisition. Caution is however required, given that this can easily be conflated with other data that are freely available as online resources; for example, the type of data that might be derived from Facebook, Twitter or other more general sources such as Google. Although many of these 'free data' resources originate from information volunteered by the public (e.g. changing of Facebook status or sending of a georeferenced Tweet) (Goodchild 2007), they are typically owned by those private organizations utilizing them in the provision of their services (Zook and Graham 2007), and as such, terms of use could easily be changed, or, the resources withdrawn from public.

Progressive moves toward Open Data are creating frameworks through which geographic data assets that have often previously not been in the public domain, or been in the public domain under more restrictive licenses, can be released for free to re-use in either commercial or non-commercial applications. Web portals are now run by various levels of administration across a number of countries of the world, providing search, links or direct downloads to data (and associated metadata) in a variety of formats. For example, national open data portals include data.gov for the US or in the UK data.gov.uk. Within countries, there are often also further local portals for administrative areas, such as a city (e.g. London: data.london.gov.uk, New York: nycopendata.socrata.com); or, for an aspect of the public sector (e.g. Policing: data.police.uk). The available formats in which data can be downloaded vary, and range in sophistication from application programming interfaces (API), through to direct download of data files in common formats (e.g. CSV). In all cases, the data are available to all users without cost and typically require a simple acknowledgement statement or signup for API access. Open Data are not necessarily restricted to attributes of places, and also include explicitly GIS data products. For example, in the UK, a number of the national mapping agency data products are now provided under open licenses (www.ordnancesurvey.co.uk/opendatadownload/products.html); or in a global context, the availability of OpenStreetMap data (www.openstreetmap.org).

Government rhetoric surrounding the release of much Open Data related to public service delivery relates to three issues: that availability will lead to improved government efficiency through the data being utilized in new innovative ways, and that this will in turn induce cost savings; second, that commercial exploitation by the private sector increases productivity and produces an economic return that is claimed back through taxation; and finally, that these data form part of a wider open governance transparency agenda to build greater public trust in politicians and other recipients of publicly generated income. The release of Open Data is however not without cost to the government; it is expensive to collate, update and distribute, and for these reasons, the returns generated by these resources may become increasingly under scrutiny. However, such evaluations are unlikely to occur over the medium term given that the various business models which do and will exploit these data are very much embryonic. It would be difficult to imagine how a government could step back from principles of openness, thus it seems reasonable to expect that Open Data resources will remain in perpetuity as part of national spatial data infrastructures (SDIs).

One of the main challenges to open data sharing is the risk of disclosing individual information; however, recent technical innovations dramatically reduce the risk of accidental/statistical disclosure. A field of research emerging from statistics and mathematics called 'Differential Privacy' ensures that additions, deletions and queries to a database do not allow a user to identify individuals (McSherry and Talwar 2007, Dwork 2008, 2009). That is, differential privacy allows an analyst to perform analytical operations on data that they do not physically own. For example, a remote server could host data and allow analytical operations upon that data, provided that such operations do not prevent privacy risks to the individuals described by the data. This model of distributed data ownership enables the potential for an entirely new model of data sharing and replication. However, the infrastructure, research and testing necessary to implement such systems in practice require significant investment. It is unclear how to finance such systems and who might bear the risk/liability for accidental data disclosures. This model of computationally mediated publishing and access to data resolves several critical issues for open GISc: the ownership of data is separated from the use of data. Sensitive data can be safely used and accessed for the purposes of replication without disclosing sensitive individual information. Finally, access to and use of data can be monitored and restricted by data owners.

For truly Open GISc, data used in analysis must be available freely to all classes of users and in perpetuity. This is essential given that there can be time lag between conducting research and publication; and that third-party researchers may revisit earlier works at a much later date. This has significant implications for storage and archiving of data related to research. For secondary data that have relatively long collection cycles such as decennial Censuses, these issues are lessened, as the storage burden is low given that their refresh rate is typically every 5 or 10 years. However, for other types of data, these issues are far more challenging; for example, high resolution satellite data could have a daily refresh rate for large parts of the world, thus creating very large storage requirements. As such, there is a necessity for the development of SDIs that could mitigate such issues, for example, enabling specific research projects to capture and document data used within online repositories (Goecks *et al*. 2010). Furthermore, efforts to improve data archiving such as the Data Documentation Initiative (www.ddialliance.org) may provide better tools for the description of both the provenance of data and its characteristics. Such services are also beginning to become available for GISc research (e.g. the UK GoGeo service http://www.gogeo.ac.uk/gogeo/). However, as argued elsewhere, access to data through such services are only the first stage of reproducibility, and that any deposits should be accompanied by appropriate metadata describing provenance, quality, credit, attribution and methods (Bechhofer *et al*. 2013).

5. Reproducible publications using workflow models

The most reproducible research requires a method of describing the process of turning spatial data into information. However, a workflow model can assist with this and is defined as a list of software instructions that produces a set of tasks required to construct a publication. These would directly link data, methods of analysis and presentation. Although, a caveat to such models is that they may not fully describe the processes by which results featured in a final presentation were created, for example,

they may not describe non-linear processes such as exploratory spatial data analysis or models tested but not used.

Examples of code and or data being required or encouraged as part of submissions to academic peer-reviewed journals are found in a variety of fields not limited to economics,[3] social simulation,[4] environment,[5] science,[6] Biostatistics[7] and computer science.[8]

There are numerous ways in which such workflows could be achieved, for example, through the construction of 'notebooks' in programming language such as Python (LeVeque 2009); however, of those potential methods available, one combination gaining increasing traction within scientific communities is the combination of R with an extension package called 'Sweave' (Leisch 2002, Koenker and Zeileis 2009). This provides R with a set of functions that enable R code to be embedded and run within a LaTeX document (see Figure 1). LaTeX (www.latex-project.org) is a document markup language that employs its own code to structure and style a text document and is notably popular among researchers whose publications make heavy use of equations given the advanced typesetting versatility. Sweave enables R commands to be embedded into the body of a LaTeX document as code snippets. When the LaTeX document is pre-processed, Sweave runs the embedded R code in chronological order, generating the specified outputs (e.g. tables, graphs, maps or diagrams) as standard LaTeX and inserting these into the document in place of the original embedded R code. The resulting document is then typeset using standard LaTeX. For those without experience of LaTeX, it should also be noted that there are alternates to Sweave such as 'odfWeave' (Kuhn 2014) which processes R code contained within documents produced in word processor friendly formats. Others have also built workflows around the R Markdown, which embeds R syntax within Markdown code, which can be rendered into a variety of formats through a library such as Pandoc.[9] Given that open data are typically available online, a LaTeX document containing R snippets can be self-contained, that is, an independent researcher could take the document and run it on their machine which

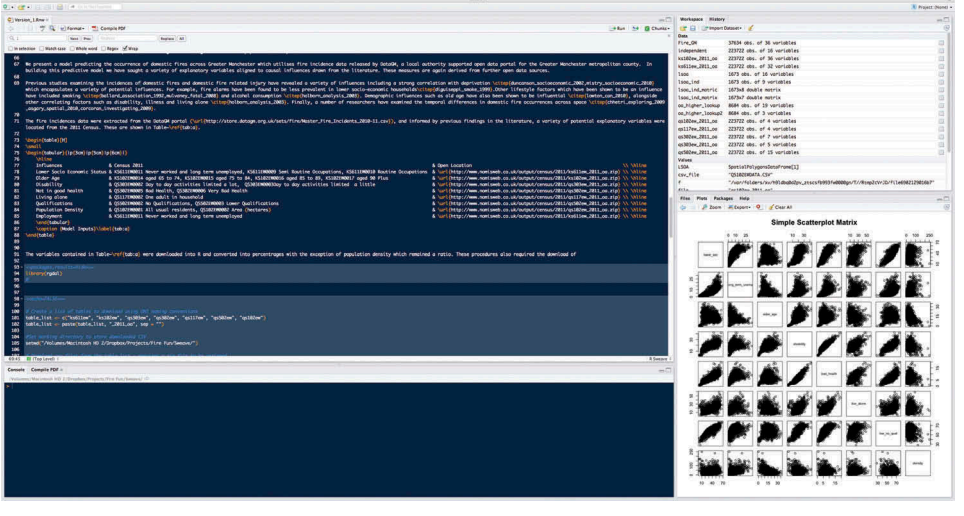

Figure 1. Latex and R code being integrated as a Sweave document within the R Studio software.

would download data, then produce an exact replica of the paper including analysis. In addition, they would also be able to adapt the R code, perhaps specifying alternate models or examining the code for errors. An alternate model is proposed by Leisch *et al*. (2011) whereby embedded analysis could be run on a server remotely as part of a third-party open source validation service, thus reducing requirements for reviewers of research to have certain software tools installed or configured. A constraint of workflows is that code may run on specific software platforms, with different versions potentially causing analysis to break, or, to return different results. Detailing such versions as part of document metadata would limit such issues; however, there is also potential for integration of emerging infrastructure such as Docker[10] which provides a common and shareable platform on which code can be run.

For a GISc workflow to be fully open, data are assumed to be scrutable and within the public domain for all users; however, GISc may also be conducted using other data sources that are closed. For example, data that are sensitive in nature, perhaps concerning attributes at the person level. As noted earlier, there are potential technical solutions to this problem, such as differential privacy, but they lack common infrastructure that would make implementation simple to manage. When GISc research is being conducted at the individual level, this makes full reproducibility difficult to embed fully within a workflow model. Within this context, developments are therefore required for a SDI that would enable geographical models to be run at an individual level, however, restricting the extent of access to the raw underlying data, thus mitigating the potential for disclosure. In the absence of such infrastructure, synthetic data might provide an alternative way to replicate results without disclosing data about real people (Abowd and Lane 2004). Such data could be generated by a 'synthesizer' that maintains the statistical properties of the input data without the actual data. For example, a variance-covariance matrix might be used to generate new data that serves as a proxy for the original data. Research into 'spatial data synthesizers', such as Quick *et al*. (2015), would thus be enormously beneficial. The US Census Bureau has started publishing synthetic individual level data based on highly sensitive administrative data from agencies like the IRS and the Social Security Administration (Bureau 2014).

A final challenge relates to the types of models that can be run within workflows. For example, a deterministic model is one that when re-run would always produce the same results, however, other applications of GISc concern stochastic problems, where results can differ each time a model is run and relate to an element of randomness in the phenomenon being modeled or technique implemented. As such, if researchers tried to reproduce such results on the basis of a workflow model, they would find these different to those presented in the previous work. However, it would be unusual when researching a stochastic problem to only interpret a single set of results in isolation, and more likely that multiple runs would be collected with the range of values presented. In these cases, the embedded code of a workflow model could include a loop, creating a collection of results from a set of analysis procedures, and then fit within this range could perhaps be examined. In such a circumstance, we would argue that it be useful for a researcher to set a 'seed' in their code to ensure that 'random' procedures are reproducible.

6. Concluding challenges for the practical implementation of open GISc

In this paper, we have argued how best practice Open GISc can be conducted within a framework where:

(1) Data should be accessible within the public domain and available to researchers. Availability might take many forms, ranging from a controlled access database to synthetic data.
(2) Software used should have open code and be scrutable.
(3) Whenever possible workflows should be public and link data, software, methods of analysis and presentation with discursive narrative
(4) The peer review process and academic publishing should require submission of a workflow model and ideally open archiving of those materials necessary for replication. These standards could be implemented through a journal's 'Instructions to Authors'.
(5) Where full reproducibility is not possible, for example, where commercial software or sensitive data are required, researchers should aim to adopt aspects of an open GISc framework attainable within their particular circumstances.

However, we recognize that adopting such a model of Open GISc is constrained by a series of SDI challenges including: overcoming software license compatibility issues, how software audit might be managed where developer numbers are limited, enabling mechanisms to manage analysis involving secure or sensitive data, and ensuring that all code (included those embedded within workflows) are well structured and documented (especially when researchers use algorithms with some inherent stochasticity).

However, we argue that despite constraints, the benefit of Open GISc for enhancing transparency of research outweigh the costs, and indeed fits within the context of wider debates about encouraging more open and accountable research practices. There are moves in this direction, for example, requirements in the UK and US that research funded through some national research agencies appear as open access. Adoption of workflow models or submission of code alongside written material within peer-reviewed journals will be a challenge, both in terms of the mechanism for submission and also how this is managed through review. However, as we discussed, there are examples of best practices appearing within other fields.

Open publishing standards, we would argue, do not dispossess researchers of their intellectual property. They simply shift the burden of reproducibility from one where the burden of reproduction is entirely on the 'reproducing' party to one where the original authors of the manuscript facilitate reproduction. We think that reproducible research is a 'higher' form of publishing, and in recognition of this, elite journals in the field should adopt clear guidelines for ensuring that work can be replicated or recognizing reproducible research.

Through this paper, we have described how Open GISc could enhance transparency of research utilizing GIS, and how this supports some of the central tenants of scientific practice. Open GISc is currently embryonic and requires a greater support from the academic community, in terms of tool development and expectations/norms in peer review. More broadly, we argue that this process could be accelerated if high quality

journals within the field of GISc began to establish mechanisms that required the submission of code and links to data as part of the article submission procedures.

Notes

1. An example includes ESRI – https://esri.github.io/
2. http://www.geoforall.org/
3. Example journals include: Empirical Economics; Journal of Applied Econometrics, Journal of Economic and Financial Modelling
4. Example journals include: the Journal of Artificial Societies and Social Simulation
5. Example journals include: Journal of Marine Science
6. Example journals include: Nature, PLoS One
7. Example journals include: Biostatistics, Biometrical Journal
8. Example journals include: the R Journal, Journal of Statistical Software
9. http://johnmacfarlane.net/pandoc/
10. https://www.docker.com/

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Alex David Singleton http://orcid.org/0000-0002-2338-2334
Seth Spielman http://orcid.org/0000-0002-5089-7632
Chris Brunsdon http://orcid.org/0000-0003-4254-1780

References

Abowd, J.M. and Lane, J., 2004. New approaches to confidentiality protection: synthetic data, remote access and research data centers. *In*: J. Domingo-Ferrer and V. Torra, eds. *Privacy in statistical databases*. Berlin: Springer, 282–289.

Baggerly, K.A. and Coombes, K.R., 2009. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 3 (4), 1309–1334. doi:10.1214/09-AOAS291

Baiocchi, G., 2007. Reproducible research in computational economics: guidelines, integrated approaches, and open source software. *Computational Economics*, 30 (1), 19–40. doi:10.1007/s10614-007-9084-4

Bechhofer, S., *et al*. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29 (2), 599–611.

Bitzer, J., Schrettl, W., and Schröder, P.J., 2007. Intrinsic motivation in open source software development. *Journal of Comparative Economics*, 35 (1), 160–169. doi:10.1016/j.jce.2006.10.001

Bivand, R., 2014. Spdep: spatial dependence: weighting schemes, statistics and models [online]. Available from: CRAN.R-project.org/package=spdep [Accessed: November 2014].

Bivand, R. and Lewin-Koh, N., 2013. Maptools: tools for reading and handling spatial objects [online]. Available from: CRAN.R-project.org/package=maptools [Accessed: November 2014].

Bivand, R., Pebesma, E.J., and Gmez-Rubio, V., 2008. *Applied spatial data analysis with R*. New York; London: Springer.

Borgman, C.L., 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63 (6), 1059–1078. doi:10.1002/asi.v63.6

Bureau, U.C., 2014. Expanding the role of synthetic data at the U.S. Census Bureau. *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, 30 (2), 117–121.
Campbell, P., 2009. Climatologists under pressure. *Nature*, 462 (7273), 545–545. doi:10.1038/462545a
Campbell, P., 2010. Closing the climategate. *Nature*, 468 (7322), 345. doi:10.1038/468345a
Cheshire, J.L.R., 2014. Spatial data visualisation with R. *In*: C. Brunsdon and A. Singleton, eds. *Geocomputation*. London: Sage, 3–20.
Dwork, C., 2008. Differential privacy: a survey of results. *In*: M. Agrawal, *et al*., eds. *Theory and applications of models of computation*. Berlin: Springer, 1–19.
Dwork, C., 2009. The differential privacy frontier. *In*: O. Reingold, ed. *Theory of cryptography*. Berlin: Springer, 496–502.
Fleck, L., 1981. *Genesis and development of a scientific fact*. Chicago: University of Chicago Press.
Gentleman, R. and Temple Lang, D., 2004. Statistical analyses and reproducible research. *Bioconductor Project Working Papers., Working Paper 2*.
Goecks, J., Nekrutenko, A., and Taylor, J., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11 (8), 1–13. doi:10.1186/gb-2010-11-8-r86
Goodchild, M. and Janelle, D.G., 2010. Toward critical spatial thinking in the social sciences and humanities. *GeoJournal*, 75, 3–13. doi:10.1007/s10708-010-9340-3
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221. doi:10.1007/s10708-007-9111-y
Herndon, T., Ash, M., and Pollin, R., 2013. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Technical report 322*, Amherst: University of Massachusetts
Jasny, B.R., *et al*. 2011. Again, and again, and again. *Science*, 334 (6060), 1225. doi:10.1126/science.334.6060.1225
Jasny, B.R., 2013. Realities of data sharing using the genome wars as case study - an historical perspective and commentary. *EPJ Data Science*, 2 (1), 1–15. doi:10.1140/epjds13
Koenker, R. and Zeileis, A., 2009. On reproducible econometric research. *Journal of Applied Econometrics*, 24 (5), 833–847. doi:10.1002/jae.1083
Kovacevic, J., 2007. How to encourage and publish reproducible research. *In*: *2007 IEEE international conference on acoustics, speech and signal processing - ICASSP '07*, 15–20 April, Honolulu, HI, IV–1273–IV–1276.
Krishnamurthy, S., 2002. Cave or community? An empirical examination of 100 mature open source projects. *First Monday*, 7 (6). doi:10.5210/fm.v7i6.960
Kuhn, M., 2014. odfWeave: Sweave processing of Open Document Format (ODF) files [online]. Available from: CRAN.R-project.org/package=odfWeave [Accessed: November 2014].
Lakhani, K. and Wolf, R., 2005. *Why hackers do what they do: understanding motivation and effort in free/open source software projects*. Cambridge: MIT Press.
Leisch, F., 2002. Sweave: dynamic generation of statistical reports using literate data analysis. *In*: W. Härdle and B. Rönz, eds. *Compstat*. Berlin: Springer, 575–580.
Leisch, F., Eugster, M., and Hothorn, T., 2011. Executable papers for the R community: the R2 platform for reproducible research. *Procedia Computer Science*, 4 (0), 618–626. doi:10.1016/j.procs.2011.04.065
LeVeque, R., 2009. Python tools for reproducible research on hyperbolic problems. *Computing in Science & Engineering*, 11 (1), 19–27. doi:10.1109/MCSE.2009.13
Longley, P.A., *et al*., 2010. *Geographic information systems and science*, Vol. 3. Hoboken, N.J: Wiley.
McMurdie, P.J. and Holmes, S., 2015. Shiny-phyloseq: web application for interactive microbiome analysis with provenance tracking. *Bioinformatics*, 31 (2), 282–283. doi:10.1093/bioinformatics/btu616
McSherry, F. and Talwar, K., 2007. Mechanism design via differential privacy. *In*: *48th Annual IEEE symposium on foundations of computer science 2007 (FOCS '07)*, October, Providence, RI, 94–103.
O'Brien, O., 2014. Open source GIS software. *In*: S.A. Brunsdon, ed. *Geocomputation*. London: Sage, 281–300.

Peng, R.D., Dominici, F., and Zeger, S.L., 2006. Reproducible epidemiologic research. *American Journal of Epidemiology*, 163 (9), 783–789. doi:10.1093/aje/kwj093

Quick, H., *et al*., 2015. Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics*, 14 (Part C), 439–451.

Raymond, E., 2001. *The cathedral and the bazaar musings on Linux and open source by an accidental revolutionary*. Farnham: O'Reilly.

Reich, M., *et al*., 2006. GenePattern 2.0. *Nature Genetics*, 38 (5), 500–501. doi:10.1038/ng0506-500

Reinhart, C.M. and Rogoff, K.S., 2010. Growth in a time of debt. *American Economic Review*, 100 (2), 573–578. doi:10.1257/aer.100.2.573

Rey, S., 2009. Show me the code: spatial analysis and open source. *Journal of Geographical Systems*, 11 (2), 191–207. doi:10.1007/s10109-009-0086-8

Rey, S., 2014. Open regional science. *The Annals of Regional Science*, 52 (3), 825–837. doi:10.1007/s00168-014-0611-7

Steiniger, S. and Bocher, E., 2009. An overview on current free and open source desktop GIS developments. *International Journal of Geographical Information Science*, 23 (10), 1345–1370. doi:10.1080/13658810802634956

Steiniger, S. and Hunter, A.J.S., 2012. The 2012 free and open source GIS software map – a guide to facilitate research, development, and adoption. *Computers, Environment and Urban Systems*, 39, 136–150. doi:10.1016/j.compenvurbsys.2012.10.003

Stodden, V., 2010. Data sharing in social science repositories: facilitating reproducible computational research. *In*: H. Wallach and J. W. Vaughan, eds. *Computational social science and the wisdom of crowds, December 10-11*, Whistler: BC.

Stodden, V.L, 2014. *implementing reproducible research*. Boca Raton, FL: CRC Press.

Sutton, T., 2010. Announcing the release of QGIS 1.6 [online]. Available from: blog.qgis.org/node/146 [Accessed: November 2014].

Tamber, P., 2000. Is scholarly publishing becoming a monopoly? *BMC Editorials*, 362 (9395), 1575–1577.

Vandewalle, P., Kovacevic, J., and Vetterli, M., 2009. Reproducible research in signal processing. *IEEE Signal Processing Magazine*, 26 (3), 37–47. doi:10.1109/MSP.2009.932122

von Krogh, G., Spaeth, S., and Lakhani, K.R., 2003. Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, 32 (7), 1217–1241. doi:10.1016/S0048-7333(03)00050-7

Warf, B. and Arias, S., 2009. *The spatial turn*. London: Routledge.

Yang, H.-L. and Lai, C.-Y., 2010. Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26 (6), 1377–1383. doi:10.1016/j.chb.2010.04.011

Zook, M. and Graham, M., 2007. The creative reconstruction of the Internet: Google and the privatization of cyberspace and DigiPlace. *Geoforum*, 38, 1322–1343. doi:10.1016/j.geoforum.2007.05.004