PopChange - An open source, reproducible research project

Nick Bearman[1][2]

[1] Centre for Spatial Demographics Research and Department of Geography and Planning, School of Environmental Sciences, University of Liverpool, Liverpool, L69 3BX, UK; [2] Clear Mapping Co Waterside House, Falmouth Road, Penryn, Cornwall, TR10 8BE, UK

Summary

PopChange is an open academic research project, developing a new approach to performing small area comparisons of British Census data from 1971 to 2011. This paper focuses on how the project was designed and implemented to be reproducible, both in terms of the implementation of the R script to generate the population adjusted grids and the accompanying resources. I justify the choice of open methods, and the benefits of selecting a range of open languages and technologies, from the academic and commercial points of view.

KEYWORDS: PopChange, census, small area comparison, R, open source, open data

1. Background

The PopChange project addresses the issue of small area estimation, and how best to compare GB Census data from 1971 to 2011 for small areas (Lloyd et al., 2016). The importance of reproducible research has been discussed extensively (e.g. Brunsdon, 2011) and in essence it is a key part of any research paper. The classic interpretation of this is the idea of a methods section which provides enough information to allow a competent researcher to replicate the findings. With more geographic research projects including the use of GIS analysis or geocomputation, the documentation and inclusion of these commands or code is more crucial than ever before. However, it is rare for a research project to include code or have space to fully document the methods in a step-by-step approach. With the push for open access research (particularly from Research Councils UK (RCUK) http://www.rcuk.ac.uk/research/openaccess/policy/) and open data alongside the development of a range of different tools (such as GitHub, RStudio and Markdown), the concept of making research truly reproducible is now becoming more feasible than ever before.

2. What is PopChange?

The ESRC-funded PopChange project (Population Change and Geographic Inequalities in the UK, 1971-2011) creates a resource to allow users to view and calculate change over time for a wide range of Census variables between 1971 - 2011 over a 1km grid for Great Britain (Lloyd et al., 2016). This includes a web based resource to access the data and perform the comparisons (see Figure 1).

Great Britain has changed radically from 1971 to 2011, with massive changes in inequality, ethnic make up, health and age structure. Currently, census data allow us to see these changes over large geographic areas, such as countries or local authorities. However, if we want to look at changes over smaller areas (e.g. a few streets, at the level of about 100 households) then this is much more difficult because while Census data are released at this level, these small areas differ in size and shape between Censuses. For example, the 1971 small area boundaries are very different to those for 2011. This project has produced population surfaces for each census year as a means of overcoming this problem.

1 nick@clearmapping.co.uk, n.bearman@liverpool.ac.uk

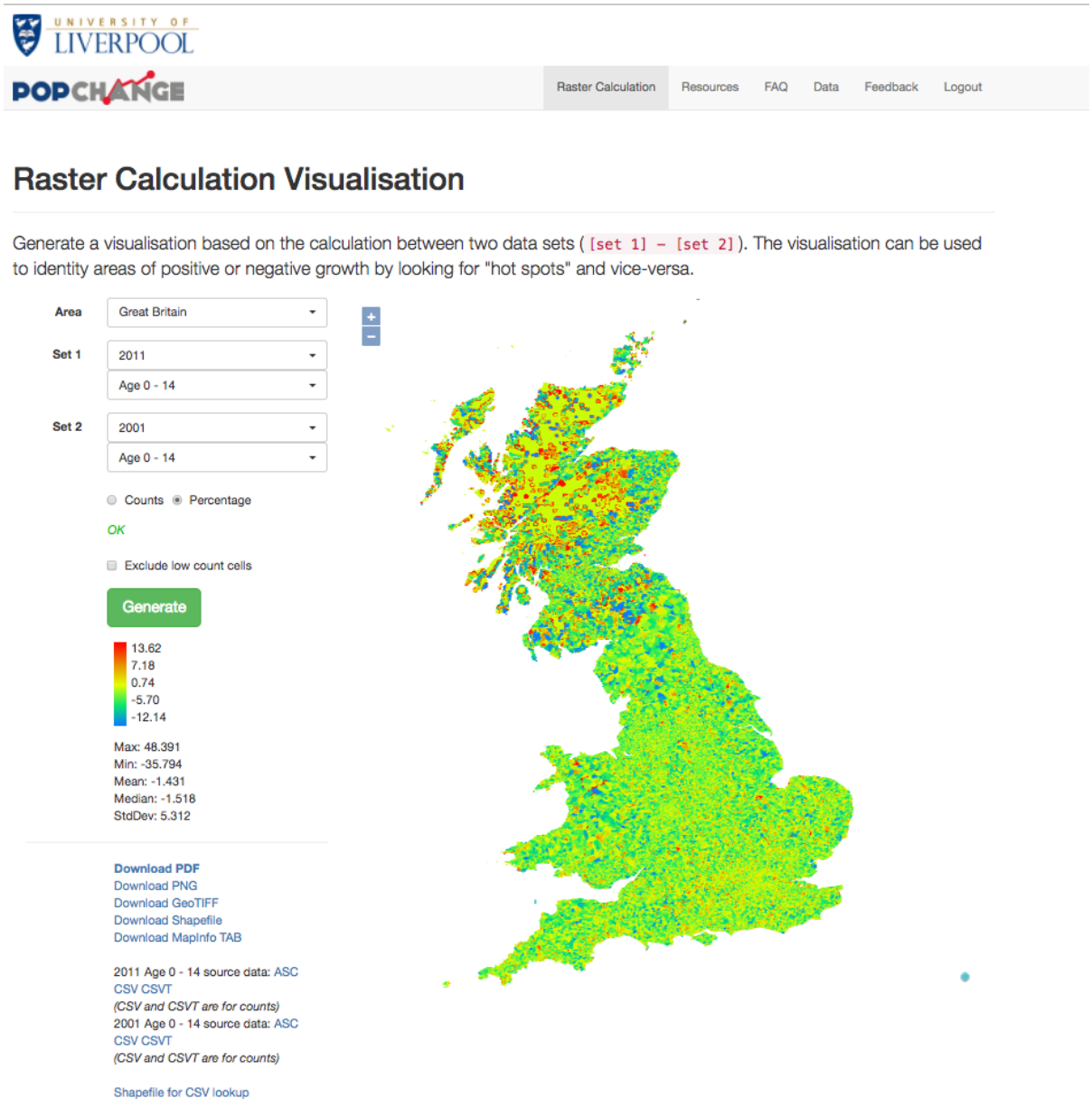

Figure 1 Screen shot of the PopChange resource

3. What is Open?

'Open' has become a real buzz word in the coding and academic communities recently. We discuss the specifics of how this has been implemented in different areas later on, and we believe that while the term has become more popular recently, the underlying principle has been core to the scientific approach for a long time.

In relation to this project, we will use the term 'open' to refer to two different things: open data and open source. Open data (both spatial and non-spatial) have been discussed extensively by others (e.g. OS Open Data, OpenStreetMap, etc.) and it is easy to forget in a spatial data heavy environment (such as GISRUK) that Open Data applies to non-spatial data as well. The Open Data Institute (theodi.org) are a leading industry body on non-spatial open data and run a range of training programs and outreach to promote the benefits of open data across a wide variety of sectors.

4. Why Open?

RCUK have an agenda of Open Access for papers to ensure that all publicly funded academic research is made available in open access journals, so there are no restrictions on who can access it.

We decided to make the PopChange project as open as possible because it allows us to show both what is possible using open Census data as well as showing how we created the tool. This is important because it will allow others to either create their own tools in a similar vein, or make changes and suggestions for the development of the PopChange resource. It also allows future projects to easily develop the resources developed within the PopChange project.

There are a number of academics who publish their code on GitHub for a variety of projects. Robin Lovelace uses GitHub to host a range of resources, including the R library stplanr, an R library providing functions and data access for transport research. Hosting the code on ROpenSci (https://github.com/ropensci/stplanr) has allowed other people to contribute to the project, enabling the code base to grow. Robin has also gone as far as writing a paper about the package stplanr that he and colleagues are in the process of submitting to the R Journal (Lovelace and Ellison, Under Review, see https://github.com/ropensci/stplanr/blob/master/vignettes/stplanr-paper.Rmd). Part of Robin's reason for an open source approach is to reduce the nature of "black boxes" in transport planning, i.e. to shed light and make available the underlying processes in transport modelling, which were previously commercial products, closed black boxes, the workings of which were understood only by the makers (Lovelace, 2016).

Chris Brunsdon also has a range of repositories on GitHub for a variety of R packages and extensions (https://github.com/chrisbrunsdon). Brunsdon has written extensively on Reproducible Research, both as a topic in itself (Brunsdon, 2016 and Singleton et al., 2016) and as a methodology within other geocomputation work, such as Geodemographic Classification (Brunsdon et al., 2014).

There are many other academics who use a reproducible research approach in their work, some using GitHub or other version control software, and some who don't. The above are simply two examples of an academic establishing themselves in their field and an established academic using open technologies and open source approaches in their work.

A range of tools have developed to allow researchers to share and develop their data or code. This has occurred in both geography and non-geography research. A recent article in Nature (Perkel, 2016) discusses how Rivers collected data on Ebola cases used GitHub (https://github.com/cmrivers/ebola) to collate her data and make it available to other researchers. She discusses why she used GitHub (ease of use and ease of access for others to use her data) as well as the benefits of making the data open and accessible, which allowed others to contribute to her dataset as well.

The open approach is not restricted to the academic area. The freelance software developer we worked with on this project (Ben Whorwood) also prefers to use an open source approach, for a variety of reasons. Firstly, from the development process point of view, open source code works as knowledge sharing, i.e. showing how you created a particular function or solved a particular problem allows others to solve that problem as well. For example, in this project Ben struggled to find any good examples of how to perform basic raster calculator functions within PostGIS and SQL. The code he wrote for this could form a good example of this for others to copy and use in their own work. In addition, it works as a learning resource, as providing the code in a navigable interface (such as GitHub) allows anyone to navigate through the code, see how it is structured and how different languages can interface with each other.

Additionally, open source code stops vendor lock in, where the user is tied to one particular vendor. This both allows the client to move vendor if she or he wishes, but also allows another person to take over the project should the initial contractor become unavailable for any reason. When the vendor is a freelancer, this is of particular benefit, as this provides more flexibility to move vendors if required. It

also promotes collaboration, both during the development phase of the project and after the project, allowing others to contribute additional features or functionality to the project.

In addition to the reasons outlined above from a freelance point of view, governments and large commercial software companies are beginning to see the benefits of the open source approach. For example, Microsoft recently had a significant change of opinion and are now actively supporting a range of open source projects, and have also included Bash (the Unix terminal / command line) within Windows (http://www.zdnet.com/article/ubuntu-and-bash-arrive-on-windows-10/)) and have recently released the .NET framework as open source (https://blogs.msdn.microsoft.com/dotnet/2014/11/12/net-core-is-open-source/), which has had a significant impact for a wide range of users. The UK Government have also begun to support and use open source formats and software in their internal departments (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/78964/Open_Source_Options_v2_0.pdf).

5. How did we make the project Open?

The growth of easy to use tools for version control (such as GitHub) have made making code available and open much easier than it has been in the past. Ease of use is a key aspect for the uptake of any technology, and the programming world is no exception to this. There are a range of different ways of making a project open source. Many academic projects make the code they use open by including it as an appendix in an academic paper, or including a zip file download on the authors web page. More recently, DOI compliant resources have been made available (such as ShareGeo, https://www.sharegeo.ac.uk/) allowing users to upload their resources and gain a permanent link that will not change.

We decided to use a version control system to allow the source code to be available and to allow others to contribute to the code and resources, as well as being able to show how the code developed over time. This also prevents us from needing a separate version control system, to allow us to role back changes and consult previous versions of the code where necessary. We decided to use GitHub to store the code and to make it publicly accessible. GitHub is an online platform for the Git version control system, allowing open source files to be hosted freely. There are a number of version control systems available, but Git and GitHub are by far the most popular and (arguably) GitHub is the easiest to use (https://blog.gitprime.com/git-didnt-beat-svn-github-did). A number of other geo researchers use GitHub, and it is widely used for data sharing and collaboration across academic research (Perkel, 2016).

Alongside the main R code to generate the grids, and the Clojure code for the web app, there are a variety of supporting codes and documents which we decided to make available through GitHub. The majority of these are written in a Git friendly format (i.e. plain text) with a focus on either R scripts or Markdown. We wanted the entire project to be as open as possible, so we made as much content as possible available.

6. Benefits and Limitations of Open

Many of the benefits of ensuring our data and code are open source stem from the desire to enable and encourage future use of the data and collaboration in further projects. Making the data and code as easy to access as possible will facilitate this and ideally encourage others to do the same. It also fulfils a range of data access requirements for the funders, which aim to ensure publicly funded research is publicly accessible.

However, there are some limitations of making the data and code open source. With both the R script and the web app code, our initial development of the code was quite informal and not necessarily suitable for publishing in a publicly available Git repository, even as a previous commit of a set of now-working code. Therefore in both the projects, our first commit of code was of an initial alpha version, with subsequent refinements through successive commits. This is a very common approach

for open source projects, both within the software community and the academic community. For the documentation, we were more open and some of the initial commits contain partially complete practicals, depending on the progress of the work at different points. While making the R script available through GitHub was quite straight forward, all of our documentation had to be written in Markdown or LaTeX, which required some transition from more commonly used Word docs. There was a learning process for all involved to adopt the new ways of working, but this was not too difficult as we were a relatively small team (three in total).

There is an argument for making the whole code development process open, highlighting the mistakes and issues we all face when writing code. This is something that could be considered for future projects, and would require all the code to be written from the first stage with this in mind. One thing we didn't implement was making the draft of the academic paper publicly available. The paper is still in process, and it could be argued that this should be made publicly accessible. Our two reasons against this were firstly that the initial draft can be quite informal (similar to the concern with the code above) and also the potential for others to take advantage of our draft paper and publish it before us.

7. Discussions and Conclusion

The PopChange project is a great example of how different elements of an academic research project can be made open and accessible. The model we have discussed works well for academic projects. In addition, from a commercial point of view, open source has a range of benefits including working as a show case project, showing to the world what our team can do.

For a project like this to be a success, it is important to discuss up front with the whole project team how open you want the final product and the development process to be. It is important to get everyone to agree to the principles underlying the project for it to be implemented successfully. As we have seen in this project, open environments for collaboration necessitate frameworks which are sometimes unfamiliar to participants, but the benefits outweigh the limitations, both in terms of short term project development and longer term academic reputation and impact.

Acknowledgements

The research on which this article was based was supported by the Economic and Social Research Council (Grant Ref. No. ES/L014769/1) and this is acknowledged gratefully. I would also like to acknowledge the other members of the PopChange project team, Chris Lloyd, Gemma Catney, Alex Singleton and Paul Williamson.

Biography

Nick Bearman is an Honorary Lecturer at University of Liverpool and Senior GIS Analyst & Course Director at Clear Mapping Co. He teaching GIS (QGIS & RStudio primarily) to a range of academics, public & private sector staff and works regularly with open source GIS for a wide range of clients.

References:

Brunsdon, C., (2011) Reproducible Research: What Can Geocomputation Achieve? *In GISRUK2011. University of Portsmouth, UK.*

Brunsdon, C., Charlton, M. and Rigby, J. (2014) An Open Source Geodemographic Classification of Small Areas In the Republic of Ireland *In GISRUK2014, University of Glasgow, UK.* http://www.gla.ac.uk/media/media_401738_en.pdf

Brunsdon, C. (2016) Quantitative methods I: Reproducible research and quantitative geography,

Progress in Human Geography 40:5 687-696, doi:10.1177/0309132515599625

Lloyd, C. D., Bearman, N., Catney, G, Singleton, A. and Williamson, P. (2016) PopChange. Liverpool: Centre for Spatial Demographics Research, University of Liverpool. https://www.liverpool.ac.uk/geography-and-planning/research/popchange/introduction/

Lovelace, R., (2016) Open source software for transport planning, Presented at UCL Centre for Advanced Spatial Analysis (CASA), 2016-11-02, London, UK. https://rpubs.com/RobinLovelace/223788

Lovelace, R., Ellison, R., (under review) stplanr: A Package for Transport Planning. *The R Journal.* https://github.com/ropensci/stplanr/blob/master/vignettes/stplanr-paper.Rmd

Perkel, 2016, Nature 538, 127–128 (06 October 2016) doi:10.1038/538127a. http://www.nature.com/news/democratic-databases-science-on-github-1.20719?WT.mc_id=TWT_NatureNews

Singleton, A.D., Spielman, S. and Brunsdon, C. (2016) Establishing a framework for Open Geographic Information science, *International Journal Of Geographical Information Science* 30:8, doi:10.1080/13658816.2015.1137579