

A Lightweight Visualization Tool for Microdata Based e-Research

Anja Le Blanc¹, Martin Turner¹, Simon Peters²

¹Manchester Computing, University of Manchester, United Kingdom.

²School of Social Sciences (Economics), University of Manchester, United Kingdom.

Email address of corresponding author: anja.leblanc@manchester.ac.uk

Abstract. One of the difficulties in processing the results of an empirical quantitative data modeling process is in post-analysis filtering or presenting the information produced. Things are a little easier if the data has a geospatial dimension: one can then use GIS software, specialized statistical modeling software, or produce "mash-ups" for web-based tools. These are all heavyweight solutions. This article presents a lightweight solution which is useful for Grid based or standalone implementation of a problem.

Introduction

Statistical applications that use social science microdata, whether cross-sectional or longitudinal (panel) studies, suffer from a lack of easy-to-use presentational and exploratory tools. Economic applications, for example, tend to result in a myriad of results tables and/or basic graphics. If such applications are Grid based, in the sense that data hosting and computation are orchestrated via the appropriate middleware, then finishing off the workflow by producing "difficult to digest" hardcopy seems to be a step backwards. However, if the application has a spatial dimension then a natural approach is to use a choropleth map style presentational and exploratory display. Tools of this type are common in the GIS (Geographical Information Systems) domain, but not necessarily easy, or cheap, to deploy.

This article presents a lightweight version of one of these tools, which was originally developed for an e-Social Science pilot demonstrator that investigated UK ethnic minority welfare. The original project, entitled Grid Enabled Microeconomic Data Analysis (Peters *et al.*, 2006,2007), Grid enabled two different microdata sources, the British Household Panel Survey (BHPS) and the 1991 Census Sample of Anonymised Records (SARs), and performed calculation of poverty measures and associated statistics using a high performance computing node. The nature of the UK Census at the time meant that the microdata's geography included the UK, its regions, and an artificial local authority area (a SARs area). The project's visualization tool allowed display and investigation of the application's results, by the geography and category of interest (ethnic minority and gender in this case), using a map interface with linked graphics or tables. This used the GeoTools open source GIS Java library (Codehaus, 2006) and was deployed via the project's web interface.

Work on the tool has moved on since the aforementioned project's conclusion to address other issues including: 1) data display, 2) tool deployment, and 3) other microdata applications. For issue 1), the original map colour scheme was improved, and the issue of producing output for formal publication was addressed. The nature of the application also

suggested enhancements to the visualization process to allow results filtering by sample size and/or statistical precision (p-value for a hypothesis test, standard error for an estimate). Sample size filtering is particularly pertinent as it alludes to the types of data disclosure restrictions imposed upon microdata by governmental and other data owners, a topic which leads into the issue of tool deployment. The original tool was deployed as part of an e-Research project that required authorised and authenticated access to the application's data sources. The tool still retains this characteristic, however, to repeat the application using 2001 UK data (the present Census currency) is not feasible as the data are only available in a secure data enclave. The code, therefore, needs taking to the remote location. The present version can be used standalone and is open source. The latter point is important, as data owners are unwilling to deploy unvetted black box code on their secure servers. The third and final issue is the ability to use the tool for different applications. This requires both mapping information and microdata, the only requirement being that these are in a format suitable for the GeoTools library. For UK microdata applications that use the Census geographies, the mapping information is available from the Edina UK borders project under Athens authentication.

This short paper discusses the present state of the tool, the details associated with the extensions discussed above and developments in progress.

Deployment and Implementation Issues

The core of the tool is based upon the GeoTools open source Java library that provides methods for the manipulation and viewing of geospatial data. As it is a Java library, associated applets or applications are platform independent. Users do not need to install it on their computers as the necessary parts are downloaded as jar files when the applet is loaded. The present implementation is fixed on GeoTools 2.0 to avoid re-writing the user interface every time GeoTools is updated.

The tool itself can be used both as a Java applet embedded in a Grid based application or as a standalone Java application, the latter being particularly useful for special environments such as secure data enclaves. To use it in a Grid based setting the user must have appropriate authentication and authorisation for the service, such as an e-certificate for the NGS (the UK's National Grid Service) or EGEE (The EU's Enabling Grids for E-Science). It is designed to sit on top of a statistical analysis that has been deployed on a Grid, and is dissociated from the middleware. This isolates development of the tool from issues (fashion, sustainability) related to the evolution of the Grid. A hypothetical implementation using the P-GRADE Portal (Sipos and Kacsuk, 2006), which can replace certain of the bespoke elements of Peters et al (2006, 2007), is presented in Figure 1 below.

Both the Grid and standalone usages may require extra security to permit use of the mapping data. This will depend upon the user's mapping file provider. For example, a user has to agree to the UK Borders license agreement for the embedded maps required for working with UK Census microdata. In practice this requires a user to have an Athens username and password.

These maps are downloaded in shape format, the supported format used in the GeoTools library. The correct geographical levels (regional maps and SARs area maps) for the currency of the data sources are readily available, however, some modifications are needed to produce a UK wide map as England, Scotland, and Wales are obtained as separate distinct mapping files. These have to be combined. At the regional level, distinguishing the regions 'South

East', 'Outer London', and 'Inner London' also required manipulation of the shape data. Once the final mapping file is available then it can be linked with an appropriately formatted data file. Most data analyses will produce flat result files which can be visualized after some filtering. This requires an application to convert them into the dbf structure required by the shape file. The combination of shape file and dbf file then become the input for the visualisation applet.

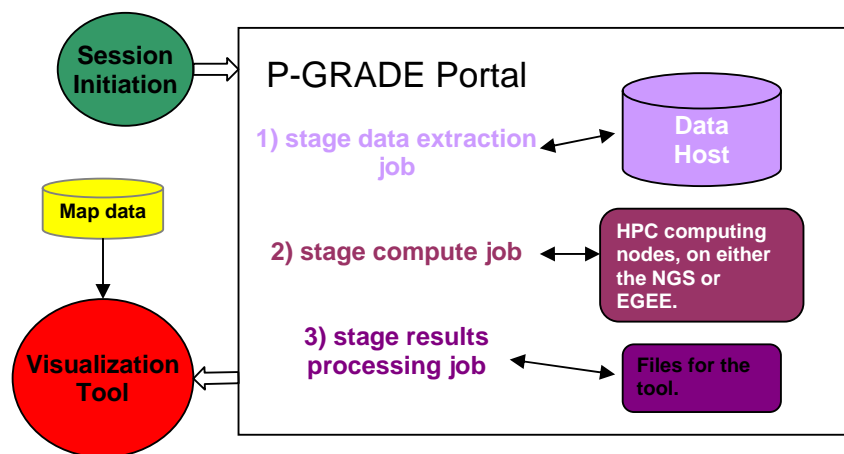


Figure 1 The Tool as an Add-on to General Middleware

User Interface Features

In certain social sciences, such as the economics discipline area, viewing the results of any microdata based modelling process in anything other than a table is still a relatively new experience. If the results of the analyses are related to specific geographies then they can be viewed using the aforementioned visualisation tool. Basic interaction (zooming, panning, etc.) within the map is possible, as well as viewing relevant statistical plots linked to a specific area of a map. The user can choose between categories pertinent to the analysis and the geographical level of the map. Our example deals with poverty measures that are produced for an ethnic group and gender category at UK regional or local area (a SARs area) geography.

Feedback from a series of peer conferences and workshops about the original visualization applet produced a number of criticisms and constructive proposals for extensions. The three main criticisms concerned: a) the chosen colour map; it was not suitable for the colour blind, b) print journal cost effectiveness; academic articles are still published in journals using greyscale by default¹ and c) extra functionality.

¹ Publishers will produce colour plates in academic journals by request. The cost for this, however, is somewhat high.

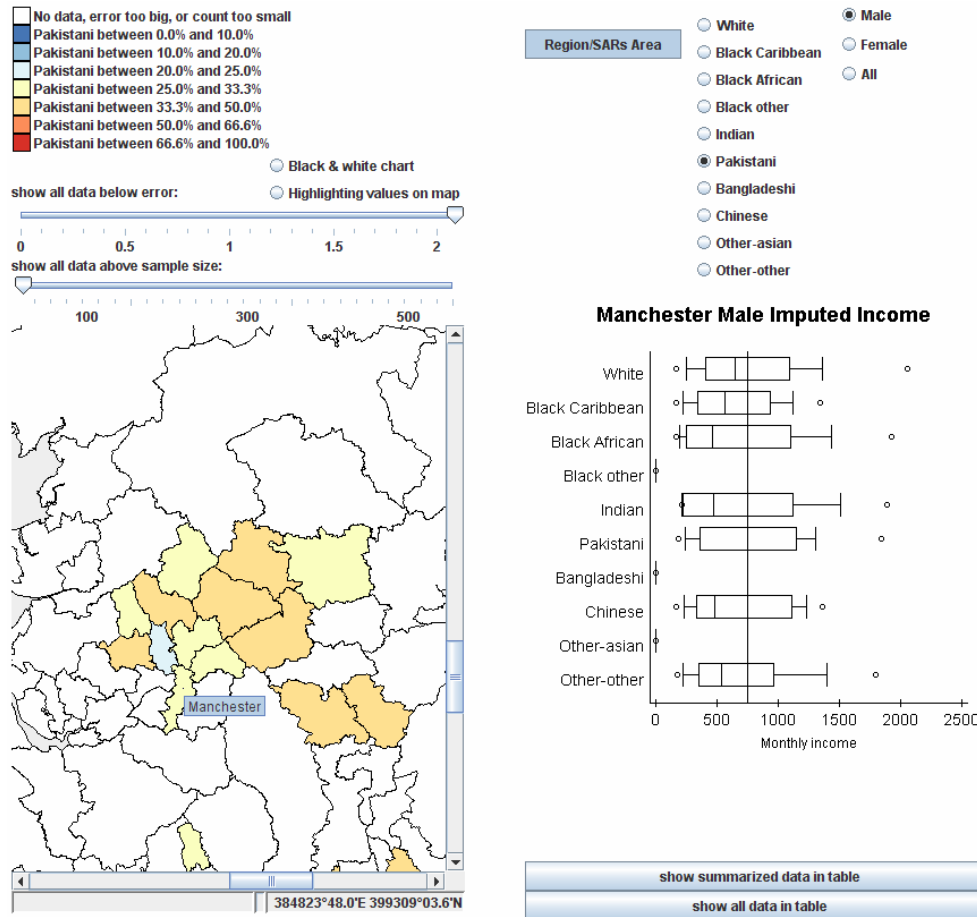


Figure 2 GEMEDA SARs area level; zoomed in with boxplot

These points were corrected in a development phase subsequent to the original project. A colour map was chosen that is both comfortable to look at for the colour seeing and distinguishable for the colour blind. The choice of colours follows ColorBrewer (2007) and Stone (2003). Additionally, an option was added to change the colour map to greyscale for printing purposes. To create a suitable grey scale for eight categories as required in this application is a problem for which there are no best practice guides in the literature. The Tufte (1983, p.154) suggestion that ten grey levels are effective for an astronomical map doesn't carry over to our examples. Our solution was to adopt a combination of five different grey levels combined with a non-obtrusive pattern for alternative levels was chosen (see Figure 3). This allowed both fast visual discrimination as well as universal printing by 'un-calibrated' printers.

Extra functionality was added to improve the capabilities for exploratory results analysis as part of the overall visualization process. Uncertainty visualization² is a key problem in data presentation, so to aid researchers in evaluating the retrieved data sliders were added to filter the data for precision as well as the sample size. With smooth control interactive evaluation can be performed by applying different filter threshold values – creating a mode of 'what-if' enquiry.

² Confidence intervals/error bars for estimate/prediction precision, p-values for hypothesis tests.

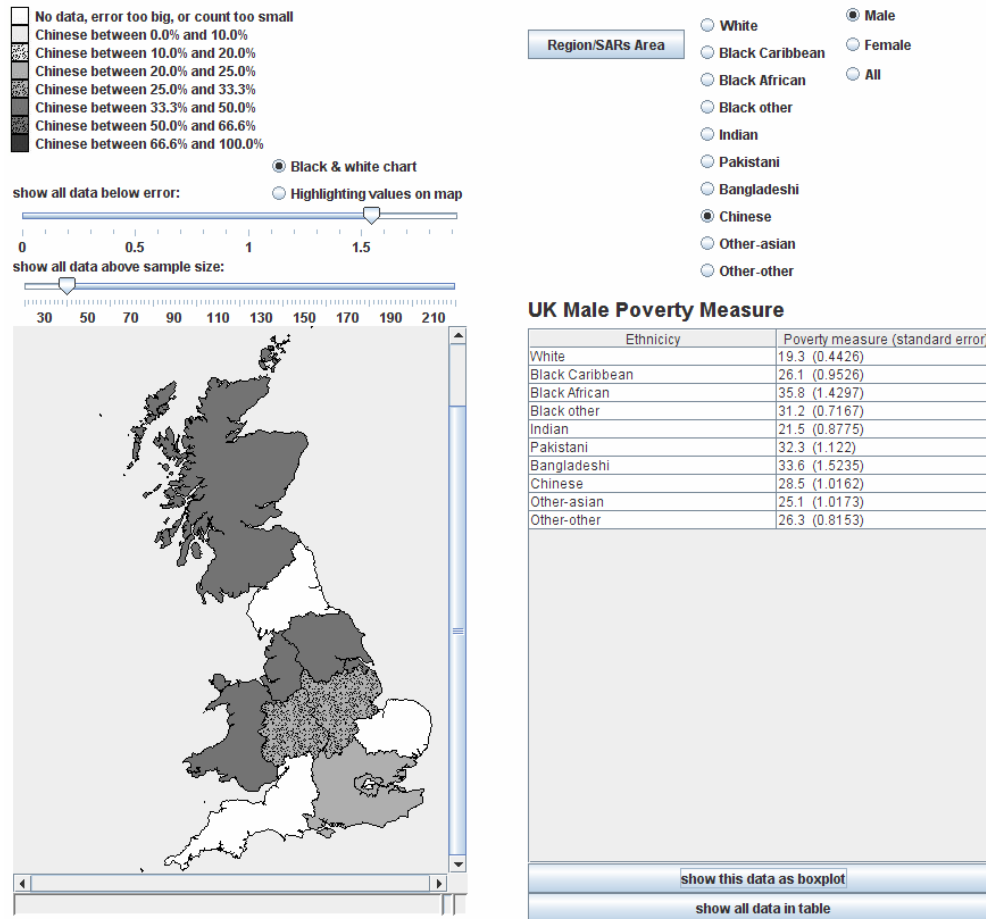


Figure 3 GEMEDA applet - black and white map with applied filters.

In geographically disaggregated microdata the information has the potential to be very sparse for categories of interest. This is particularly acute for the ethnic minority groups in our example, presenting the possibility that the global look and then zoom method of investigation could result in small regions being missed. To overcome this deficiency and make it easier to locate all the areas where data is available; an option was added to highlight in a striking colour all possible areas. After applying some zoom functions the user can switch back to the normal colour map representation. The original statistic of interest and its measure of precision (the poverty measure and its standard error in our example) can still be viewed by examining the underlying data in the form of a table. This allows cross-category comparison for a chosen geographic area.

Concluding Comments

This short article has presented ongoing developments of a visualization tool originally designed for presenting the results of a Grid based statistical analysis of UK microdata. The tool is self-contained and opensource and its development can proceed without having to engage directly with developments in either the core GIS Java library used or with middleware. It is designed to be lightweight, requiring only standard file formats for both maps and associated data as input. It can be deployed as a standalone application or as an applet to sit on top of a Grid based service. The design is suited to a variety of microdata

based statistical or econometric analysis results presentations³, and has the potential to be useful for macro models that report results at regional or lower levels of geography. Further developments will include enhancements to its functionality, however, it is not intended to be a spatial statistical modeling program (e.g. GeoDa of Anselin et al, 2004), a "mash-up" generator (e.g. the GeoVUE project, Hudson-Smith, 2006) or a heavyweight GIS.

Acknowledgments

Research supported by the National Centre for e-Social Science Hub.

We have benefited from discussions with Mashuda Glencross, Diego Gutierrez, Caroline Larboulette, Johnathan Roberts and Jamsheed Shorish.

Mapping data was provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and the ED-LINE Consortium. Copyright statement: "This work is based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and the ED-LINE Consortium."

References

- Anselin, L., I. Syabri, Y. Kho (2006), 'GeoDa: An Introduction to Spatial Data Analysis', *Geographical Analysis*, 38, pp.5-22.
- Codehaus (2006), *GeoTools The Open Source Java GIS Toolkit*, <http://geotools.codehaus.org/>.
- ColorBrewer (2007), <http://www.colorbrewer.org>
- Hudson-Smith, A. (2006), 'GeoVUE: Visualising Data via the Grid', <http://www.ncess.ac.uk/research/nodes/GeoVUE/presentations/20060920-hudsonSmith-GeoVUE.pdf>
- Peters, S., K. Clark, P. Ekin, A. Le Blanc, S. Pickles (2007), 'Grid Enabling Empirical Economics: A Microdata Application', *Computational Economics*, vol. 27.
- Peters, S., P. Ekin, A. Le Blanc, K. Clark, S. Pickles (2006), 'Grid Enabled Data Fusion for Calculating Poverty Measures', *Proceedings of the UK e-Science All Hands Meeting 2006*, pp.536-533
- Sipos, G. and P. Kacsuk (2006), 'Multi-Grid, Multi-User Workflows in the P-GRADE Portal', *Journal of Grid Computing*, 3, pp.221-238.
- Stone, M.C. (2003), *A Field Guide to Digital Color*, AK Peters
- Tufte, E.R. (1983), *The Visual Display of Quantitative Information*, Graphics Press.

³ Space precludes presenting further examples. Some should be available in the supporting presentation material.