Panemalia: visualising longitudinal datasets at the Australian Data Archive

Kevin Pulo[1], Ben Evans[1], Deborah Mitchell[2], Steven McEachern[2]
[1]ANU Supercomputer Facility, Australian National University, Canberra, Australia. Ben.Evans@anu.edu.au
[2]Australian Data Archive, Canberra, Australia, Deborah.Mitchell@anu.edu.au

ABSTRACT

Longitudinal surveys are a very rich form of social science data, often containing a wealth of as-yet untapped hidden knowledge. However, such datasets are typically examined using analytic techniques and simple graphs. We believe that much better can be done in the analysis and exploration of such fertile datasets. Panemalia is the application of an advanced visualisation technique to longitudinal survey data. It is a highly interactive DHTML application, integrated with the data repository at ADA, is accessible by non-IT savvy social science users, and supports the requirements of data familiarisation, exploration and quality assurance.

INTRODUCTION

Social science, as a discipline, is very good at collecting, managing, archiving, processing and analysing social data in a variety of forms. However, *data visualisation* is one area that has traditionally not been fully capitalised by the social sciences. When datasets are analysed by social science researchers, the most common approaches are to use purely analytic techniques, such as cross-tabulations or regressions, or to use simple graphs, such as scatterplots or pie charts. This is despite social science datasets being extremely large and rich, and containing a wealth of “hidden information”.

Visualisation is an excellent way of gaining insights into data that may not otherwise be apparent. However, datasets that are very large, very rich, or both, can present problems for simple visualisation techniques which often scale poorly as the number of data points or variables increases dramatically. This necessitates the use of advanced, and often sophisticated, data visualisation methods so as to allow people to see such large datasets in an effective way. These techniques also often involve a level of interactivity and navigation, to deal with the fact that there is more data than pixels on the screen. This interactivity additionally permits users to have an intimate experience with the data, as they explore its intricacies.

The Panemalia tool applies advanced techniques to the problem of visualising longitudinal survey data. This data has exceptionally high dimensionality (many hundreds or thousands of variables), over many thousands of individuals, over multiple collection waves. The visualisation was required to be easy to use and accessible to ordinary users, which means that a direct and conceptually straightforward visualisation technique must be used, with support for interactive exploration and view manipulation.

The tool has been integrated with the ADAs repository of longitudinal data. For these longitudinal datasets, Panemalia is available as an additional option to all users who have access to download the dataset. A separate “Visualisation” tab lists all the Panemalia-enabled datasets that are available, as well as news and usage information for the tool itself.

METHODOLOGY

The visualisation technique used is known as Parallel Coordinate Plots. In this technique, variable axes are laid out in parallel (as opposed to orthogonally in regular 2D and 3D graphs). Data points are then plotted as lines connecting the respective values on the axes. This technique allows the use of many variables (by simply lining up many axes), and is a direct visualisation because the values of each line can be easily read from each axis. However, the patterns that are visible now depends on how the axes are arranged, and this heightens the need for the user to be able to interactively rearrange the data being visualised.

Panemalia is targeted at three main use cases.

1. Initial familiarisation. Users unacquainted with a particular dataset should be able to use the tool to get an initial “feel” for the data and its characteristics, perhaps in anticipation of performing analysis such as regressions.
2. Finding information. Researchers should be able to use the tool to find evidence which supports hypotheses, or to find facts or information previously unknown. This may be by users explicitly looking for interesting features or supporting evidence, or it may be in the form of “serendipitous discovery”.
3. Data cleaning. Archivists who are preparing datasets for submission into the archive would benefit from being able to use the tool to perform quality assurance on the data, for example, finding transcription errors (as gross outliers), logical inconsistencies of responses between questions or waves, and so on.

The tool itself is easy and intuitive to use, in much the same way as many other modern advanced web-based applications. As established earlier, a high degree of interactivity is required for users to be able to adequately explore the data, and this is supported by best-practice user-interface design. Wherever possible, direct interaction metaphors are used, for example, dragging objects to move them, and drawing selection boxes directly on data of interest. Changes in the user's view are animated wherever possible, as this helps to preserve the user's “mental map” of the data.

Panemalia is a DHTML application that runs in the user's browser. The main framework is jQuery (as with the rest of the ADA website). The jQueryUI toolkit is used for the user-interface controls in the left-side panel, and the main visualisation itself is created using the raphaelJS toolkit, which uses the underlying browser SVG support. This requires a reasonably recent web browser, however, it considerably eases the development of advanced features such as curved lines and smooth animations. Before a dataset can be used in Panemalia, it must be pre-processed from the original Stata data file into an SQLite file. The data is generally read-only (with only occasional updates by archivists), and this permits good performance for many concurrent reads of subsections of the data, without the complicated overhead of a full relational database. jQuery AJAX calls are used to load variables into the visualisation on-demand as the user selects them in the user interface, with the data returned in JSON format by the Ruby on Rails web server backend.

Figure 1 shows an example screenshot of Panemalia, visualising work/study history data from the Negotiating the Life Course dataset. A selection box has been used to colour the female respondents in red, while the males are blue.

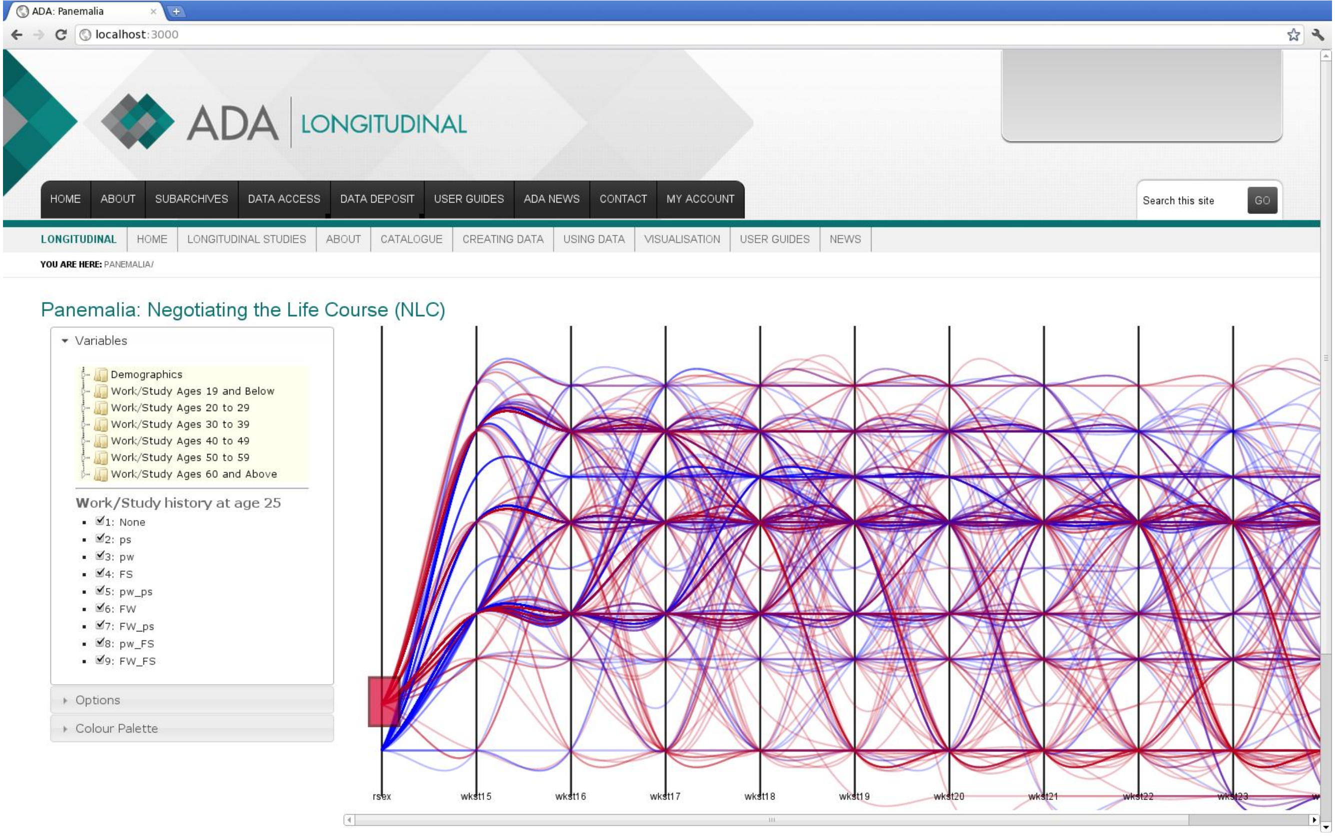

Figure 1: Panemalia example screenshot

Conclusion

Panemalia aims to enhance the ability of social scientists to visualise longitudinal survey data, permitting data familiarisation, exploration and quality assurance. Parallel Coordinate Plots are used to visualise this high-dimensional survey data in a direct and scalable way. This is achieved through the use of a highly interactive web application that is tightly integrated with the ADA website.

About the Author(s)

Kevin Pulo is an Academic Consultant and Systems Programmer at the ANU Supercomputer Facility. He works in the areas of information and data analysis and visualisation, parallel application programming, and HPC user environments. Kevin.Pulo@anu.edu.au

Ben Evans is the Head of the ANU Supercomputer Facility at the Australian National University. He leads projects in HPC and Data-Intensive analysis, working with the partners of NCI and the research sector.

Deborah Mitchell is the Director of the Australian Data Archive.

Steven McEachern is the National Manager of the Australian Data Archive.