Panemalia: Visualising Longitudinal Datasets at the Australian Data Archive

Kevin Pulo
kevin.pulo@anu.edu.au

Australian National University Supercomputing Facility (ANUSF), National Computational Infrastructure National Facility (NCI NF), Canberra, Australia

2011-11-07
Longitudinal datasets

- Respondents answer survey/questionnaire
  - Mailout, phone, etc
- Follows the same respondents over time
  - Collection “waves”
- \( \Rightarrow \) Very rich form of data
- Excellent scope for advanced and sophisticated visualisation techniques
Panemalia

- Visualisation tool for longitudinal datasets
- Easy to use
  - Integrated with the ADA website
  - Interactively manipulate view of the data
  - Visualisation method presents the data in a simple and direct way
Use cases

- Initial familiarisation (all users)
- Finding information (researchers)
  - Targeted (supporting evidence), or
  - Serendipitous
- Data cleaning (archivists)
  - Problems tend to be visually apparent
Longitudinal visualisation

- Longitudinal datasets contain:
  - Thousands of respondents, answering
  - Hundreds of questions, over
  - Multiple waves
Longitudinal visualisation

- Longitudinal datasets contain:
  - Thousands of respondents, answering
  - Hundreds of questions, over
  - Multiple waves

- The challenge:

  How to visualise this data?
Mapping data to space

- Graphs map variables *directly* to 2D/3D space
Mapping data to space

- Graphs map variables *directly* to 2D/3D space

2 variables:
\[ x, y \]

Variable 2

Variable 1
Mapping data to space

- Graphs map variables *directly* to 2D/3D space
Mapping data to space

- Graphs map variables *directly* to 2D/3D space
Mapping data to space

- Graphs map variables *directly* to 2D/3D space
Mapping data to space

- Graphs map variables *directly* to 2D/3D space
Mapping data to space

- Graphs map variables *directly* to 2D/3D space
• But longitudinal datasets have hundreds of variables (at least)
But longitudinal datasets have hundreds of variables (at least)

Cannot possibly map this directly to hundred-dimensional space
Visualising multi-dimensional data

• But longitudinal datasets have hundreds of variables (at least)
• Cannot possibly map this directly to hundred-dimensional space
• Need some way of mapping this high-dimensional data to low-dimensional space
• But longitudinal datasets have hundreds of variables (at least)
• Cannot possibly map this directly to hundred-dimensional space
• Need some way of mapping this high-dimensional data to low-dimensional space

Parallel Coordinate Plots
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

• Dimensions (axes) are laid out in parallel
• Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Parallel Coordinate Plots

- Dimensions (axes) are laid out in parallel
- Use lines to join variable values
Advantages

- Allows many dimensions
- Straightforward mapping
- Human perceptual system is good at seeing connectedness and patterns
Challenge: Ambiguity

Common values cause ambiguity
Challenge: Ambiguity

Common values cause ambiguity
Solution: (a) Curves
Challenge: Ambiguity

Common values cause ambiguity
Solution: (a) Curves
Challenge: Ambiguity

Common values cause ambiguity
Solution: (a) Curves       (b) Spreading
Challenge: Axis parameters

View affected by axis ordering, sign, scaling, translation
Challenge: Axis parameters

View affected by axis ordering, sign, scaling, translation
Challenge: Axis parameters

View affected by axis ordering, sign, scaling, translation
Challenge: Axis parameters

View affected by axis ordering, sign, scaling, translation
Challenge: Axis parameters

View affected by axis ordering, sign, scaling, translation
Challenge: Axis parameters

View affected by axis ordering, sign, scaling, translation
Challenge: Axis parameters

View affected by axis ordering, sign, scaling, translation

Solution: Interactive view manipulation
Pilot implementation

- Based on “parvis” InfoVis research software
- Java application
  - Swing + Java2D canvas
- Issues highlighted:
  - Web integration: possible, but not optimal
  - Data storage: flat text files not scalable
  - Interactivity: poor responsiveness, usability
- Useful to confirm the use cases and benefits of longitudinal visualisation
Pilot implementation
Panemalia technology

- **Frontend**
  - DHTML (HTML5 + Javascript + CSS)
  - jQuery + jQueryUI
  - raphaelJS (SVG or VML)
  - Requires modern browser
    - (But not Firefox on Linux)

- **Backend**
  - Ruby on Rails
  - SQLite
  - R

- **Queries and results using AJAX + JSON**
Negotiating the Life Course (NLC)

- 4 waves (unbalanced), 1000–3000 respondents
- Examine aspects relating to labour force participation
  - Work and study history
    - Retrospectively from age 15 (up to age 62)
    - Full-time, part-time, none
Work/study: interleaved by year
Work/study: interleaved by year

Panemalia: Negotiating the Life Course (NLC)
Work/study: interleaved by year
Work/study: interleaved by year
Work/study: interleaved by year
Work/study: interleaved by year

Panemalia: Negotiating the Life Course (NLC)
Work/study: interleaved by year

Panemalia: Negotiating the Life Course (NLC)
Work/study: combined by year
Work/study: combined by year

Panemalia: Negotiating the Life Course (NLC)
Work/study: combined by age
Work/study: combined by age
Work/study: combined by age

Panemalia: Negotiating the Life Course (NLC)

Variables:
- Demographics
- Respondent sex
- Work history by year
- Study history by year
- Work/study history by year
- Age 19 and below
- Age 20 to 29
- Age 30 to 39
- Age 40 to 49
- Age 50 to 59
- Age 60 and above

Respondent sex
- 1: Male
- 2: Female

Options
- Colour Palette

Funding Partners
- Australian Consortium for National Collaborative Research
- ANU
- The University of Queensland
- The University of Western Australia

Collaboration Partners
Work/study: combined by age

Panemalia: Negotiating the Life Course (NLC)

Variables
- Demographics
  - Respondent sex
    - 1: Male
    - 2: Female
- Work history by year
- Study history by year
- Work/study history by year
- Age 19 and below
- Age 20 to 29
- Age 30 to 39
- Age 40 to 49
- Age 50 to 59
- Age 60 and above

Options
- Colour Palette
Work/study: combined by age
Work/study: combined by age
Work/study: combined by age
Moving forwards

• Improved interactive performance
  - WebGL
  - Web workers (multithreading)
  - Remote desktop to ADA Virtual Lab VM

• State and session management
  - “How did I get to this view?”
  - “I had a great view 2 minutes ago, now it’s gone.”
  - Sessions spanning many days and locations

• Data cleaning
  - Allow sandboxed raw dataset upload, auto-convert and private visualisation
  - Standalone Panemalia in a desktop VM image