Code-based, open-source software for teaching interactive data visualisation

Shan-I Lee, BSc (Hons)
Supervisor: Paul Murrell

Department of Statistics
The University of Auckland

November 16, 2017

Problem

Tukey (1965, p. 25)

Today, software and hardware together provide far more powerful factories than most statisticians realize, factories that many of today's most able young people find exciting and worth learning about

Problem

Tukey (1965, p. 25)

Today, software and hardware together provide far more powerful factories than most statisticians realize, factories that many of today's most able young people find exciting and worth learning about

- How does interactivity benefit data analysis?
- Which interactive techniques are 'worth learning'?
- Which code-based, open-source software to use?

Method

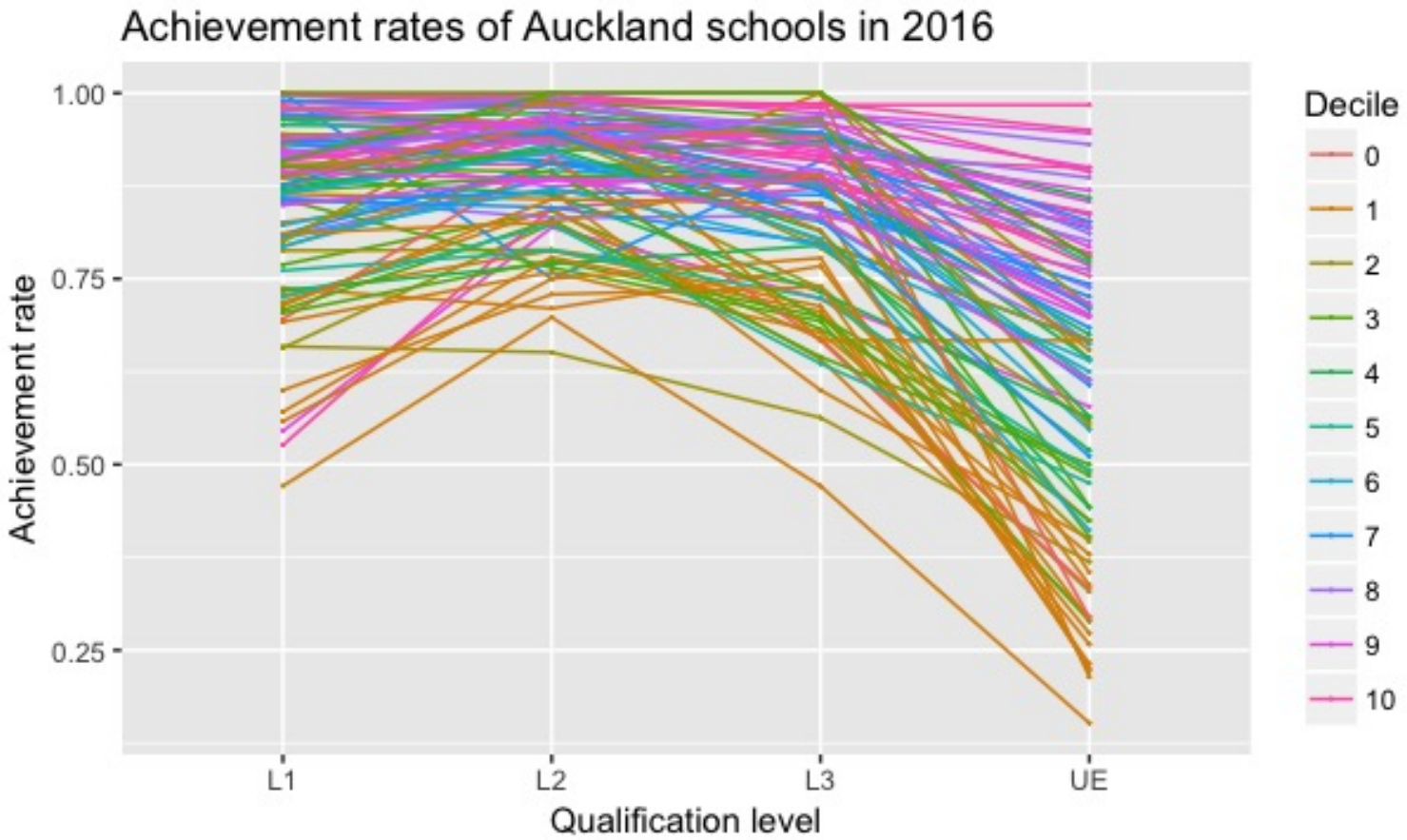
- Literature review of interactive techniques.
  - Interactive data visualisation using **GGobi** graphical user interface (Cook and Swayne, 2007)
- Survey of current code-based, open-source software.
- Application to exploratory data analysis of 2016 National Certificate Educational Achievement (NCEA) results.
  - Explore how interactive techniques further insight into data.

Findings

- Key interactive techniques that enrich data analysis:
  - Linked brushing
  - Identification
  - Scaling
  - Subset selection
  - Tours
- A focal set of **R** packages for applying interactive data visualisation: **plotly**, **crosstalk** & **shiny**.
  - Ease of installation and application
  - Coverage of interactive techniques
- The benefits of interactivity justify the effort of teaching interactive tools.

Leveraging static plots

Parallel coordinates plot (PCP)



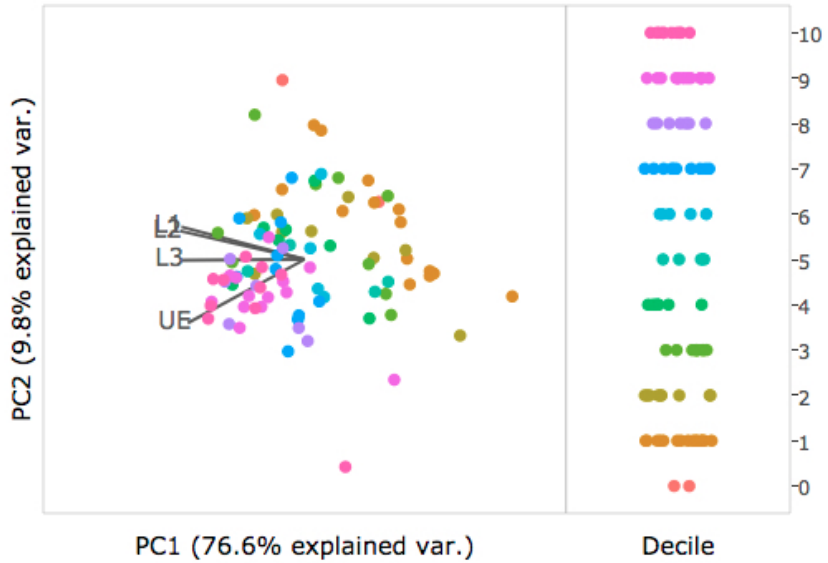
Leveraging static plots

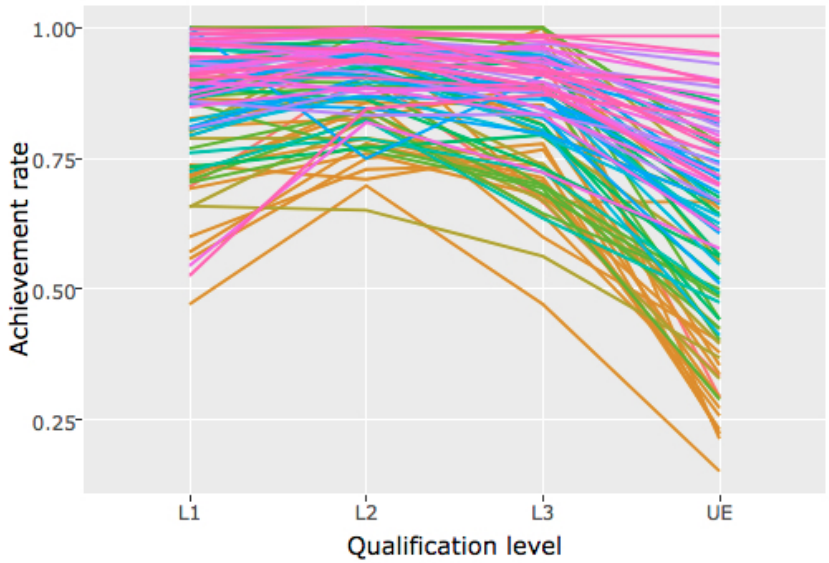
Parallel coordinates plot (PCP)

- **Linked brushing** and **identification** allow fast querying of unusual patterns, groups and/or individuals.
- Interactive **scaling** to compare and explore the different patterns revealed.
- **Subset selection** via filtering views alleviates issues with overplotting.

Relating multiple views

Principal components plot and PCP





Relating multiple views

Principal components plot and PCP

- Insights into multivariate data structures gained from individual static plots are extended by **linked brushing**.
- Applying interactive data visualisation encourages further exploration of the data.
  - Questions are quickly addressed and more questions arise from probing the data with interactive techniques.
- Awareness of the strengths and weaknesses of different software allows for efficient application of interactive techniques.

A focal set of software

Coverage of interactive techniques by **shiny**, **plotly** and **crosstalk**.

Package	Active R session	Tooltip Identification	Scaling	Subset selection	Linked brushing (except lines)	Animation (for tours)
Shiny	Yes			Analysis & filtering views	Aggregate brush possible	Yes
Plotly		Yes	Zoom in or out	Filtering views only		Yes
Crosstalk				Filtering views only	Easiest for 1-to-1	Yes

Conclusion

- Interactive techniques benefit data analysis.
  - Insights beyond static plots.
  - Relate multiple views.
  - Further exploration of structures.

Conclusion

- Interactive techniques benefit data analysis.
  - Insights beyond static plots.
  - Relate multiple views.
  - Further exploration of structures.
- The **R** packages **shiny**, **plotly** and **crosstalk** enable interactive data visualisation with code-based, open-source software.

Conclusion

- Interactive techniques benefit data analysis.
  - Insights beyond static plots.
  - Relate multiple views.
  - Further exploration of structures.
- The **R** packages **shiny**, **plotly** and **crosstalk** enable interactive data visualisation with code-based, open-source software.
- The benefits of applying interactive techniques to data analysis warrant teaching interactive data visualisation to future statisticians.

References I

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application Framework for R*. R package version 1.0.3.

Cheng, J. (2017). *crosstalk: Inter-Widget Interactivity for HTML Widgets*. R package version 1.0.1.

Cook, D. and Swayne, D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis With R and GGobi*. Springer Publishing Company, Incorporated, 1st edition.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

References II

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., and Despouy, P. (2017). *plotly: Create Interactive Web Graphics via 'plotly.js'*. R package version 4.7.0.

Tukey, J. W. (1965). The technical tools of statistics. *The American Statistician*, 19(2):23–28.