MONASH University

ETC3250/5250: Introduction to Machine Learning

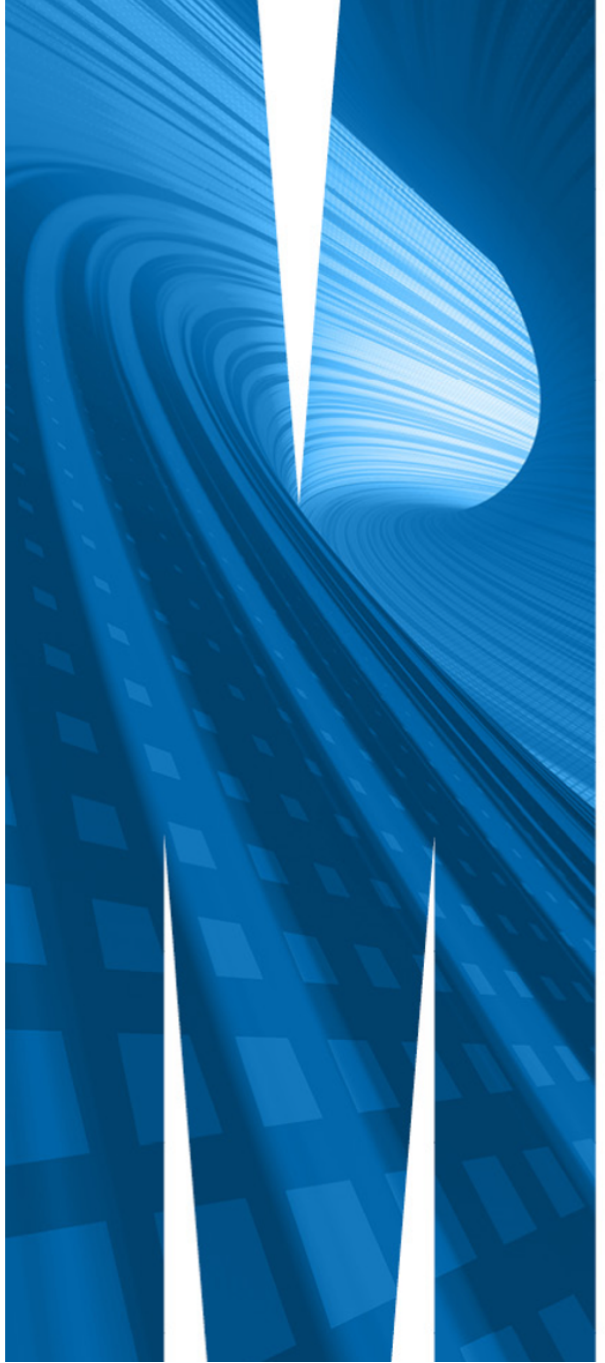
Visualisation of multivariate data

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

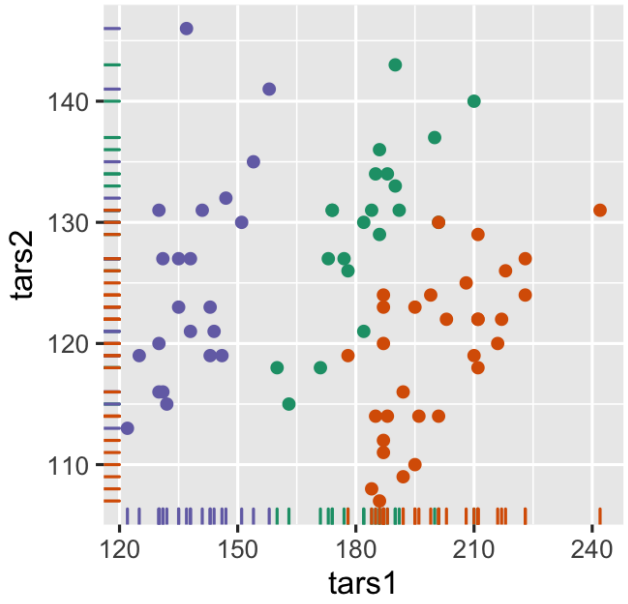
ETC3250.Clayton-x@monash.edu

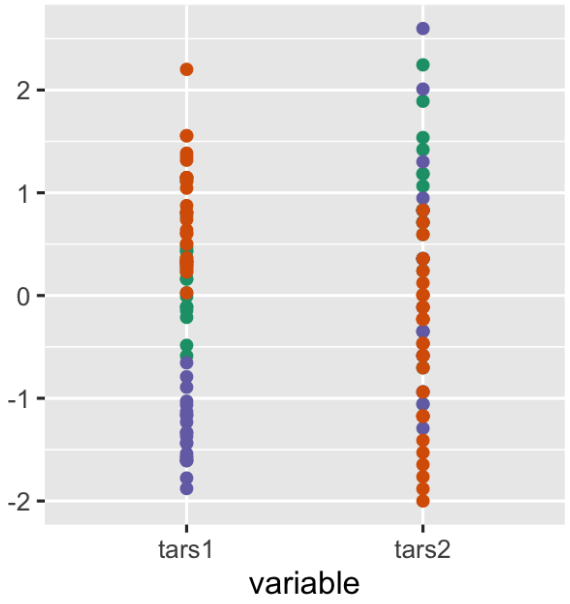
Week 5b

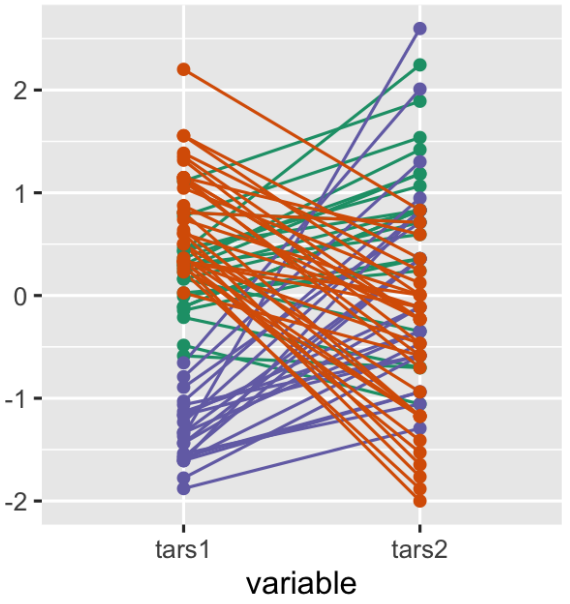


Parallel coordinate plots

- Scatterplots use orthogonal axes, and are thus limited to two variables on the page.
- Turning the axes parallel allows for many more variables to be displayed together.
- Lines connecting the points show associations between variables.

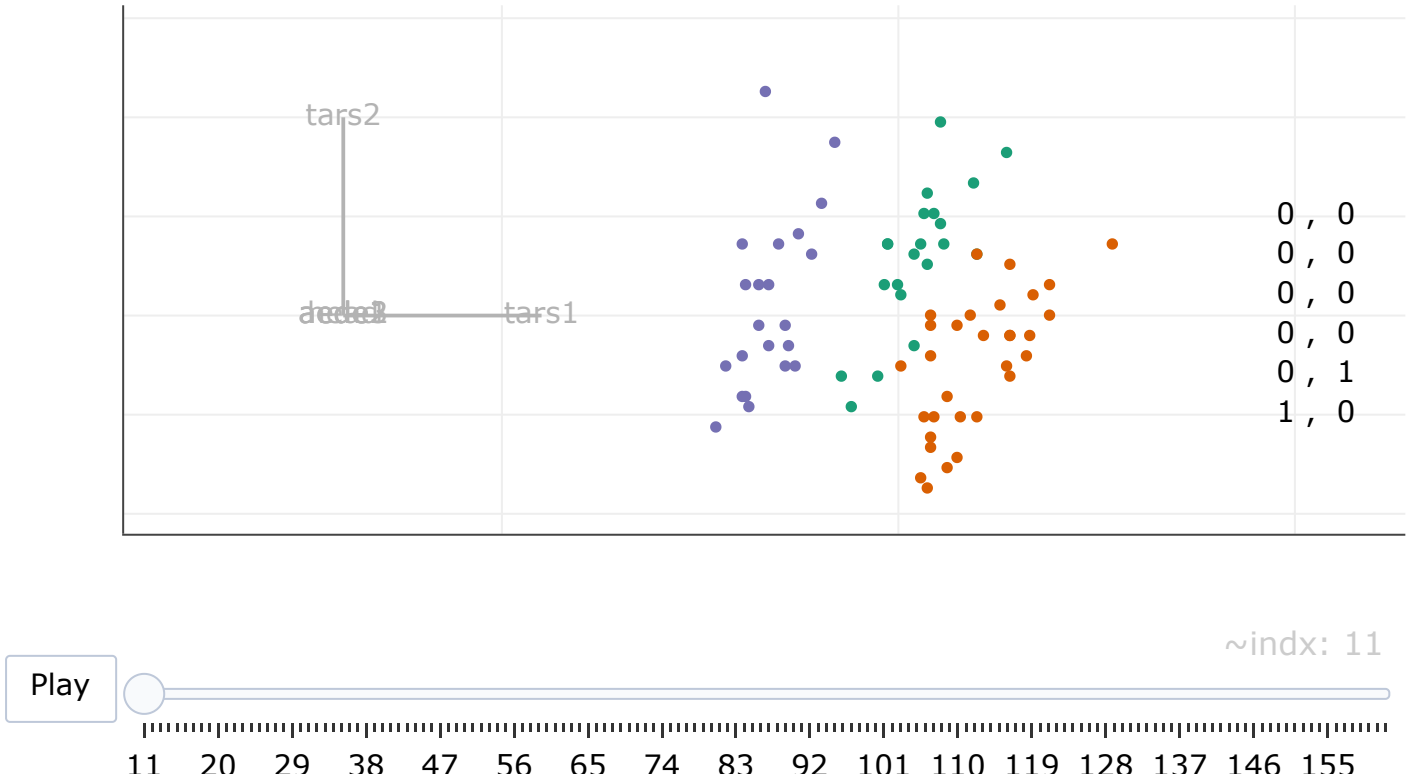






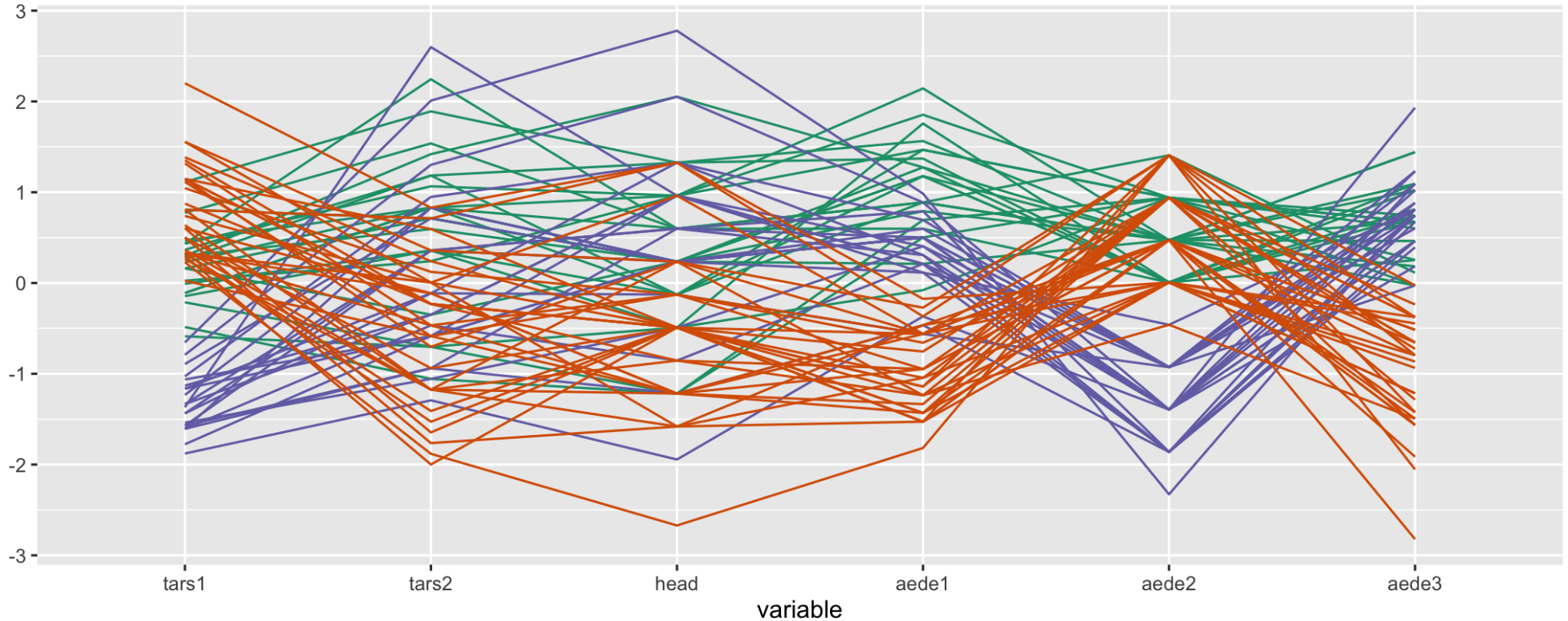
Comparison with tours

Compare the tour of the flea data, with three clusters:



Comparison with tours

With the parallel coordinate plot:



Note: the trend in lines, that there seem to be three patterns of lines.

How to read parallel coordinate plots

- A set of points in p dimensional space map to a set of lines in p parallel axes
- The pattern among and between the lines indicate structure in high-dimensions
  - Groups of lines trending together indicate clustering
  - Single lines trending differently to other indicate outliers
  - Between pairs of axes intersecting lines indicate strong negative association, and parallel lines indicate strong positive association

Points in Euclidean space ≠ lines in parallel coordinates

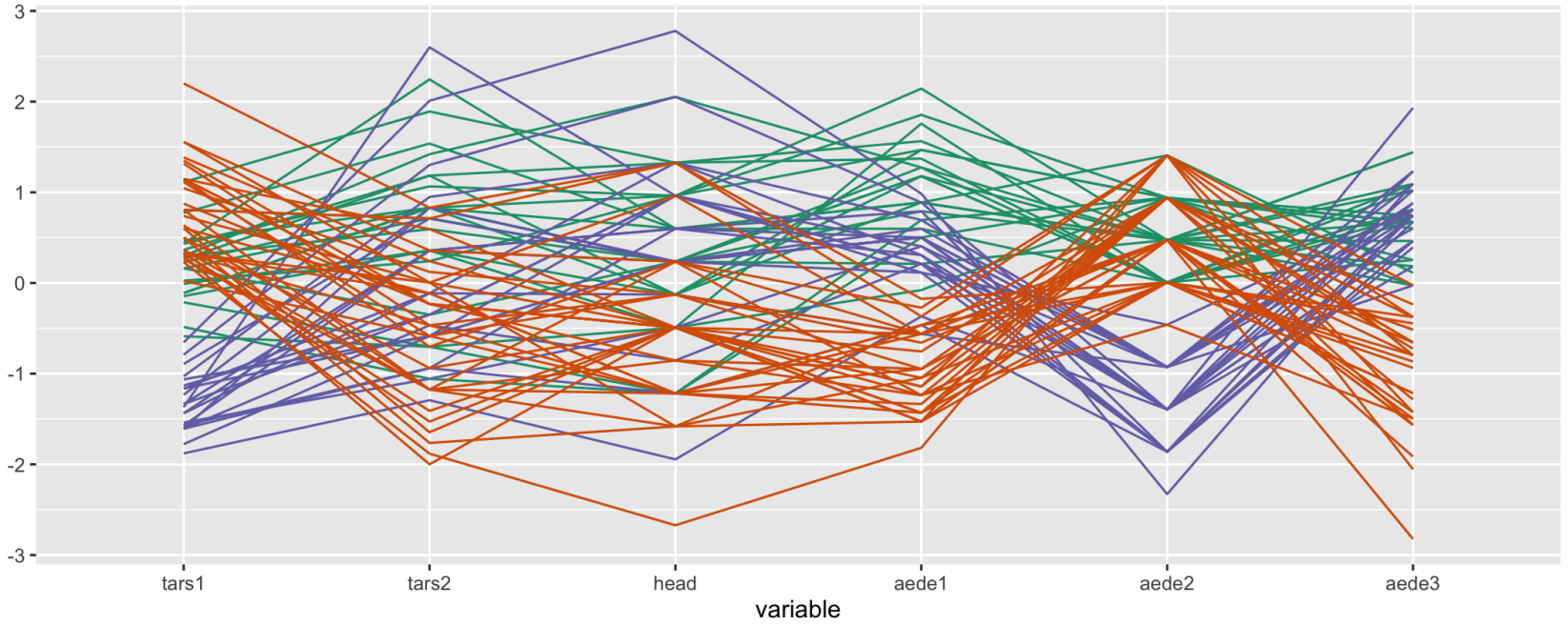
Parallel coordinate plots - controls

Details that need to be controlled:

- Order of axes can affect perception of structure. Placing axes next to each other emphasizes that association
- Variables need to be on a similar scale, and may need to be standardised

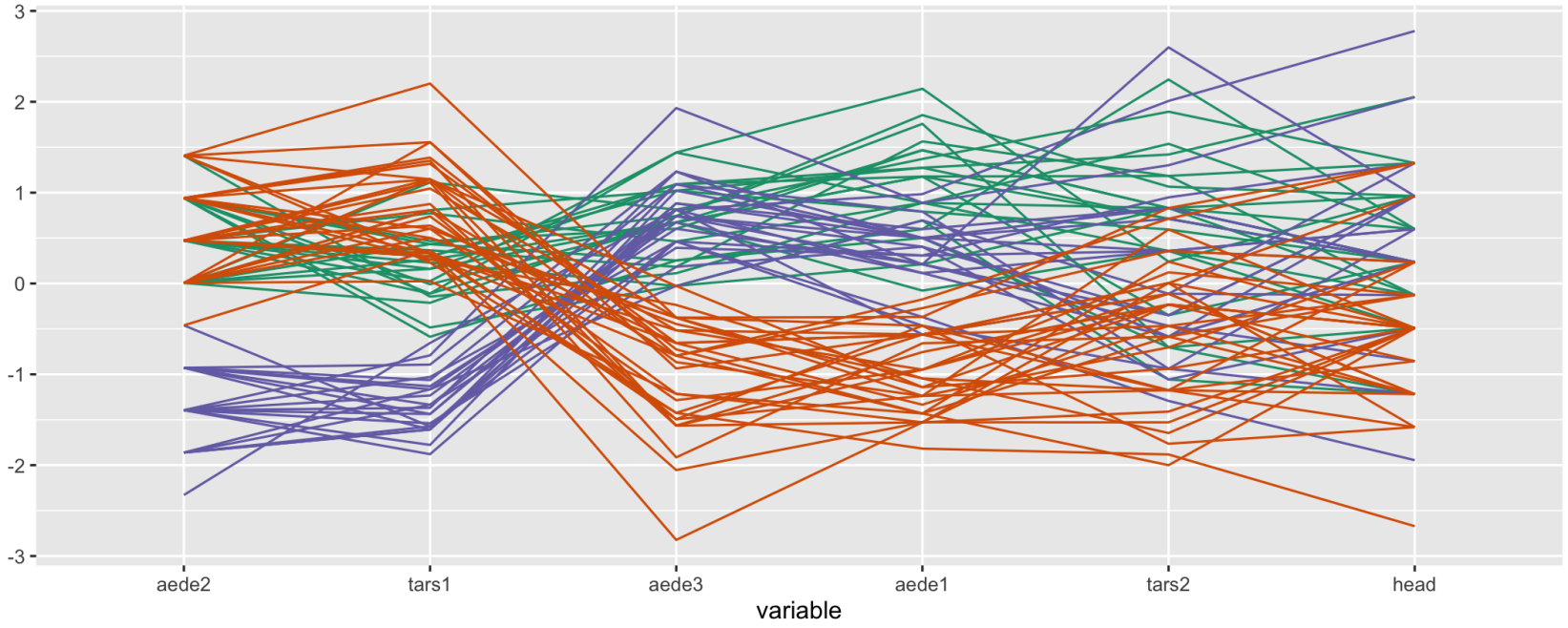
Parallel coordinate plots - controls

Unordered, default standardised scale.



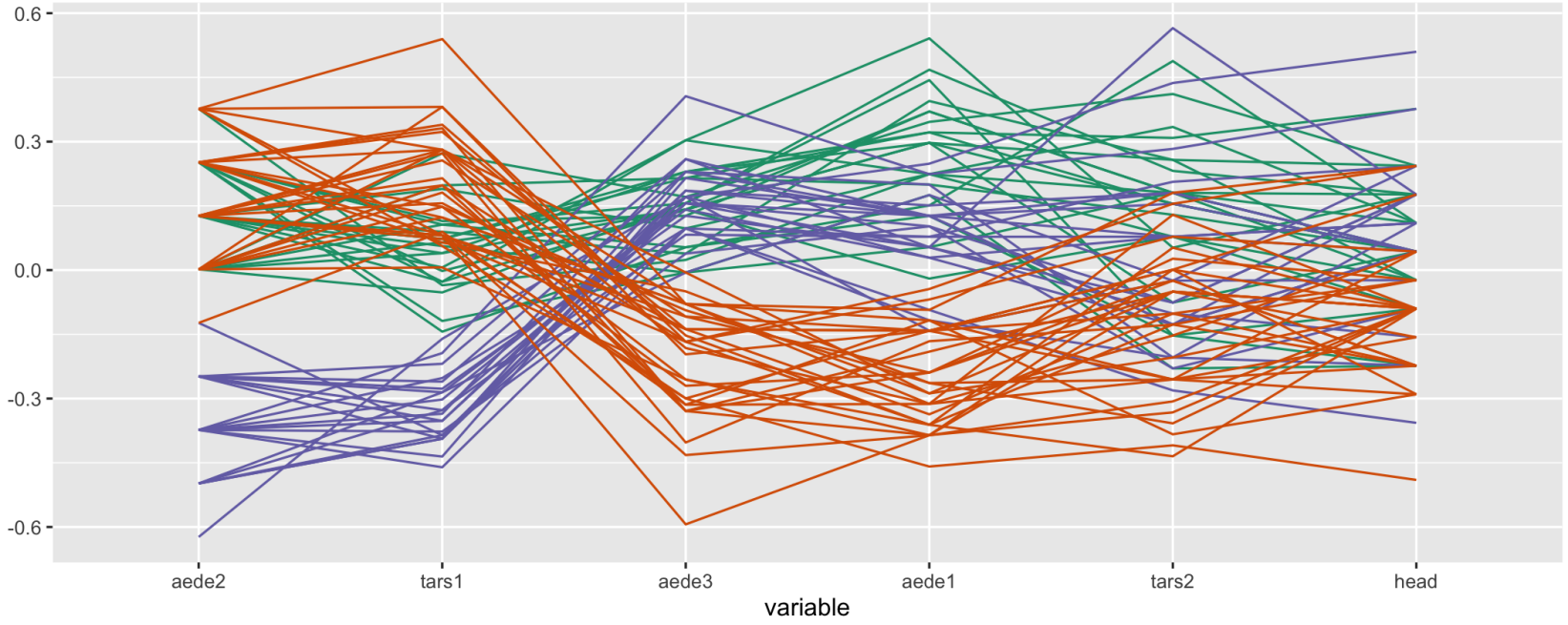
Parallel coordinate plots - controls

Ordered, default standardised scale.



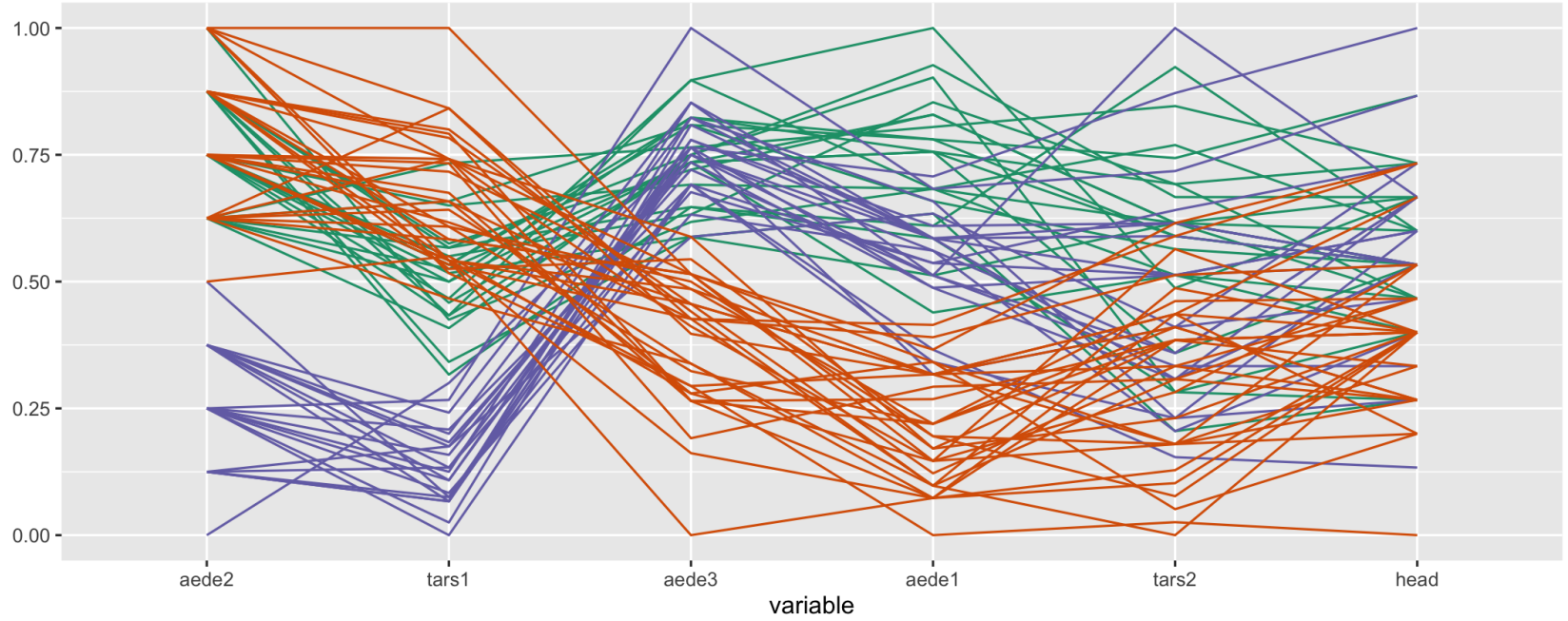
Parallel coordinate plots - controls

Ordered, centered standardised scale.



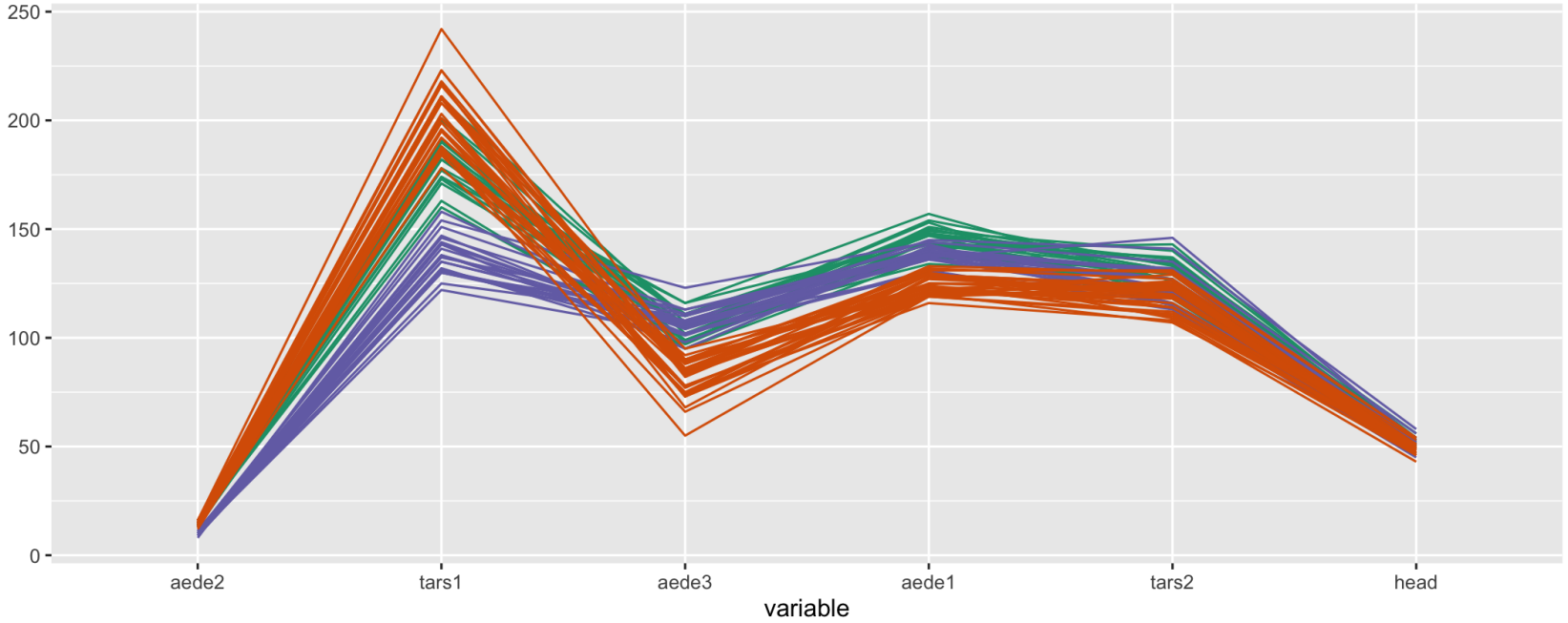
Parallel coordinate plots - controls

Ordered, scaled univariately to 0-1.



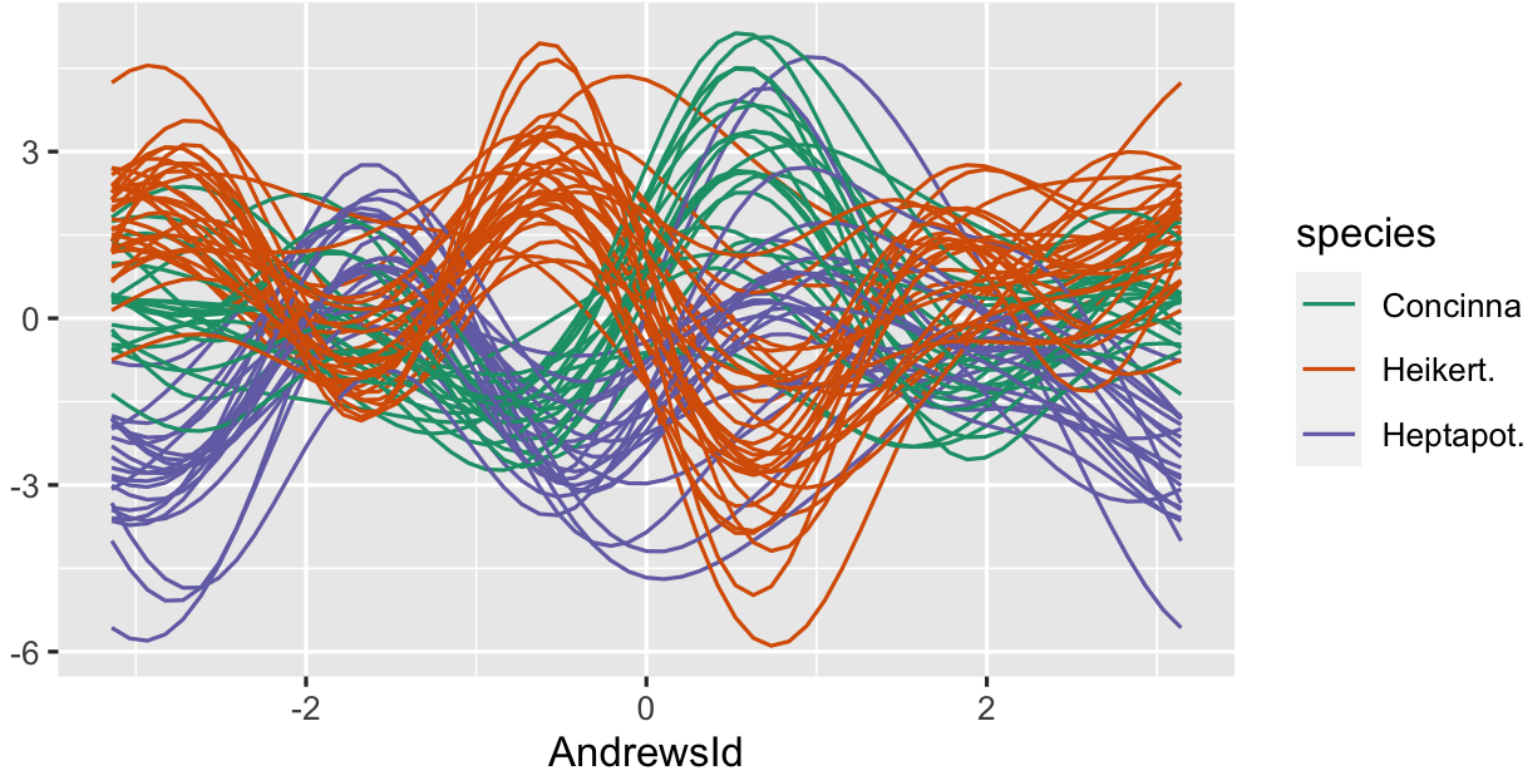
Parallel coordinate plots - controls

Ordered, scaled globally to 0-1.



Andrews curves

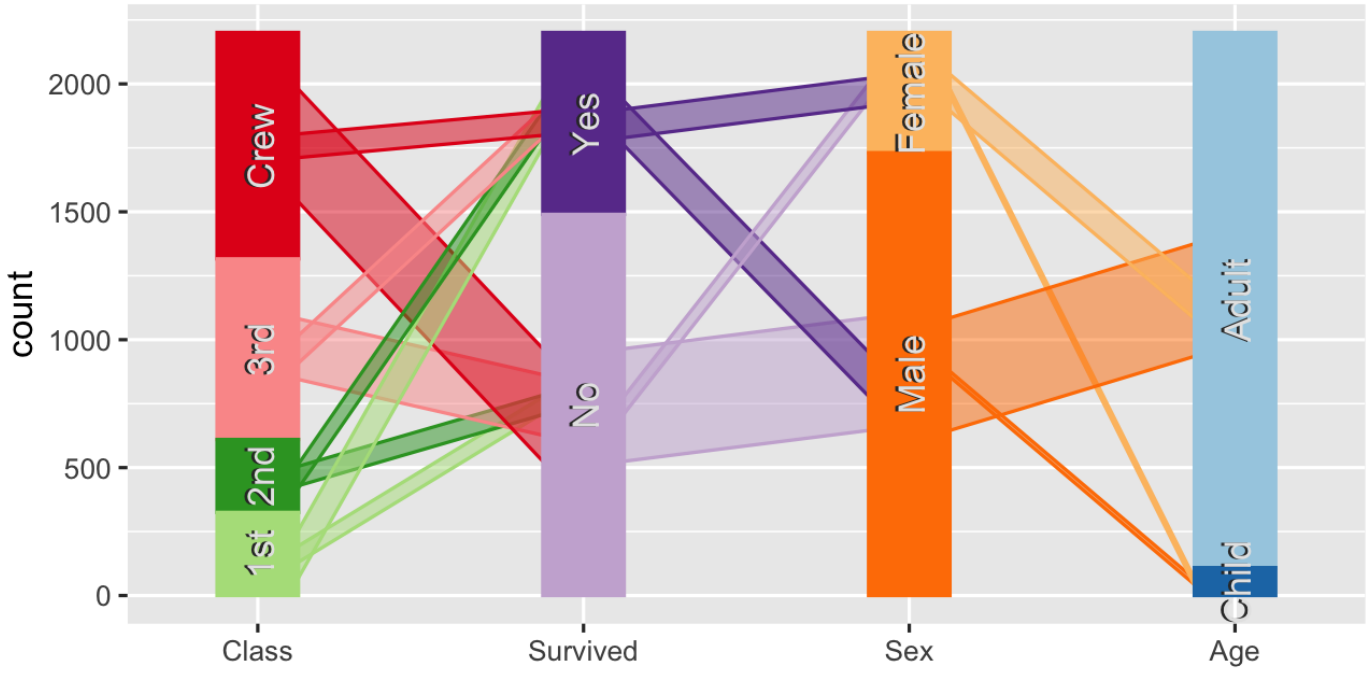
For the p variables $(x_1, \ldots, x_p)$ make a fourier transform $f(x) = \frac{x_1}{\sqrt{2}} + \ldots$.



Preceded the tour, but like a tour ($d = 1$) on a page, but algorithm is not space-filling.

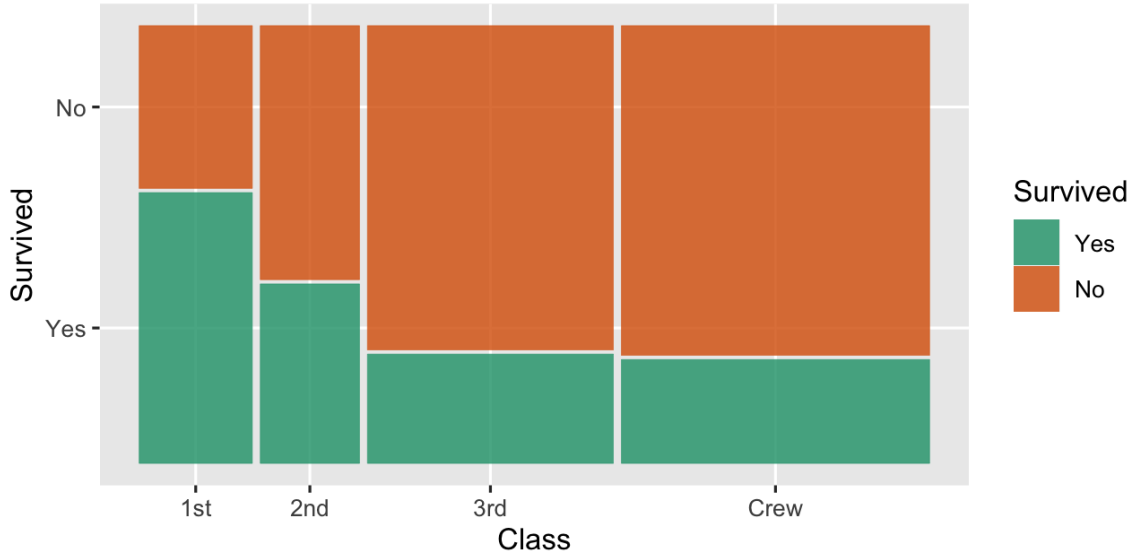
Categorical variables

Hammock plots are a variation of parallel coordinate plots for categorical variables. They show the flow of groups between stacked barcharts, providing information about the association between multiple categorical variables.



Categorical variables

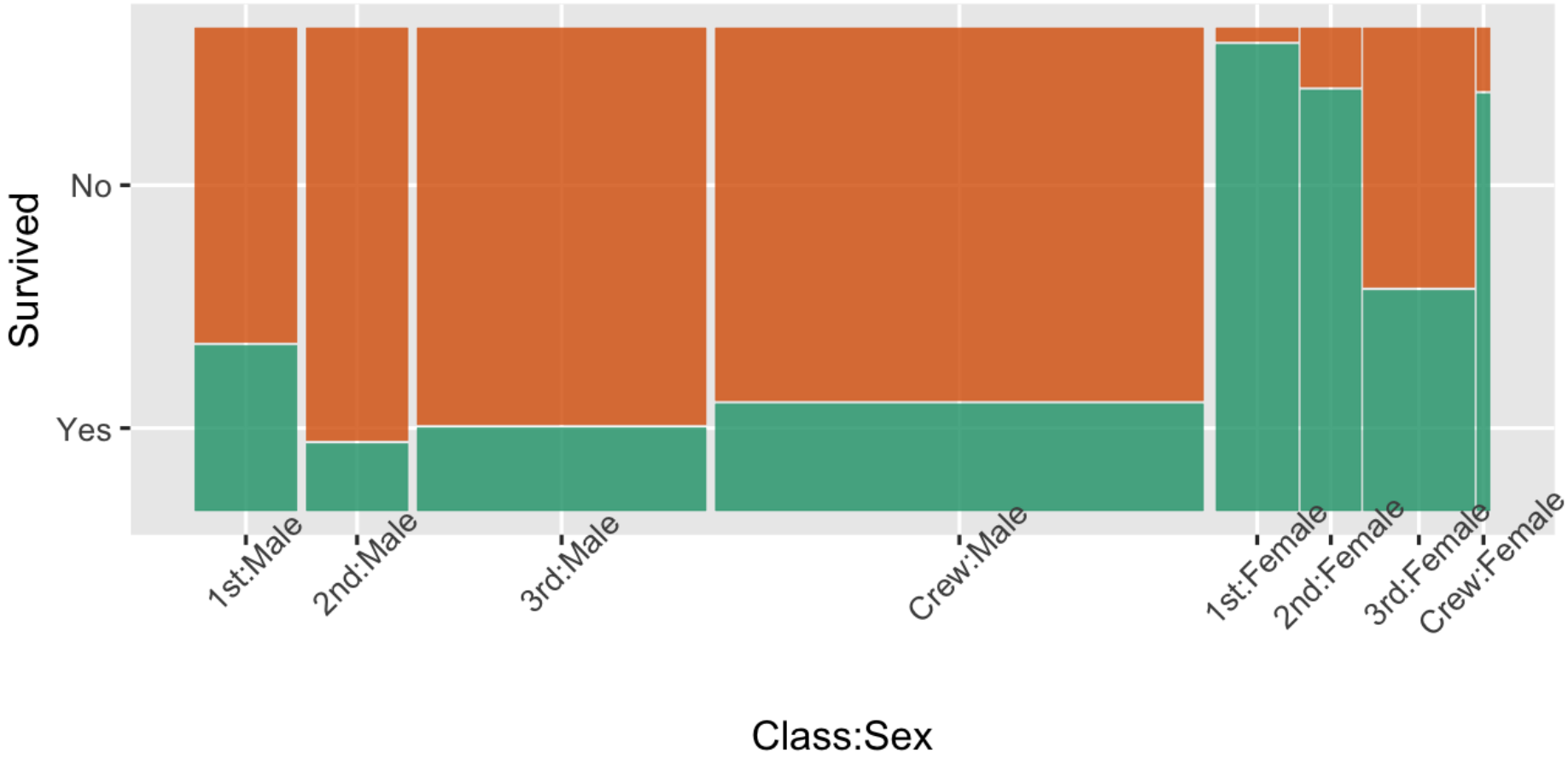
Mosaic plots partition the axes sequentially on the categorical variables.



For this data, `titanic`, there is a response variable, "Survived" and good practice would have this variable mapped to fill colour, because we are interested in the proportion change in response across the predictor categories.

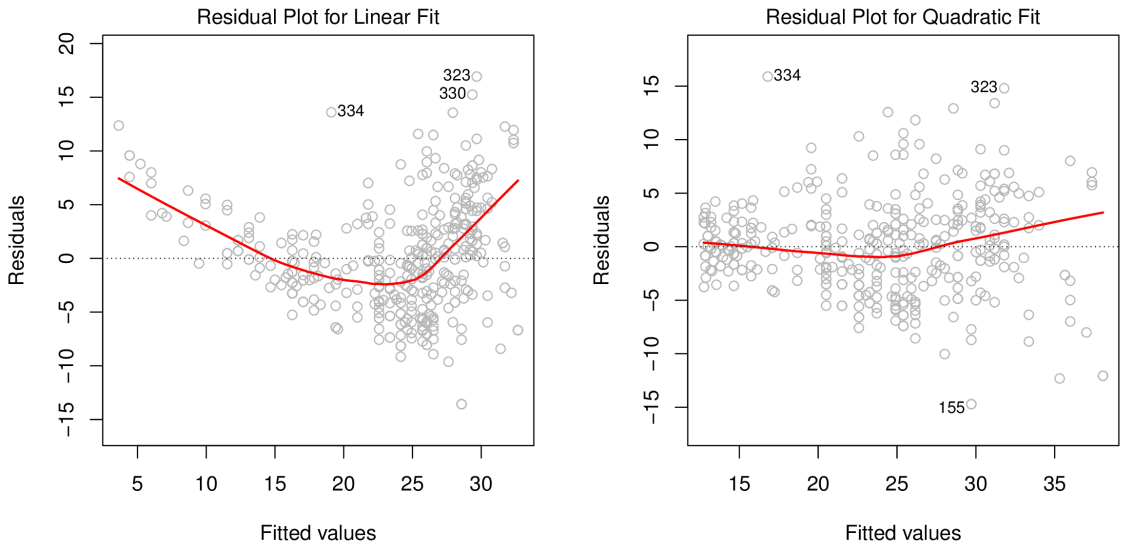
Categorical variables

Mosaic plots can handle multiple categorical variables.



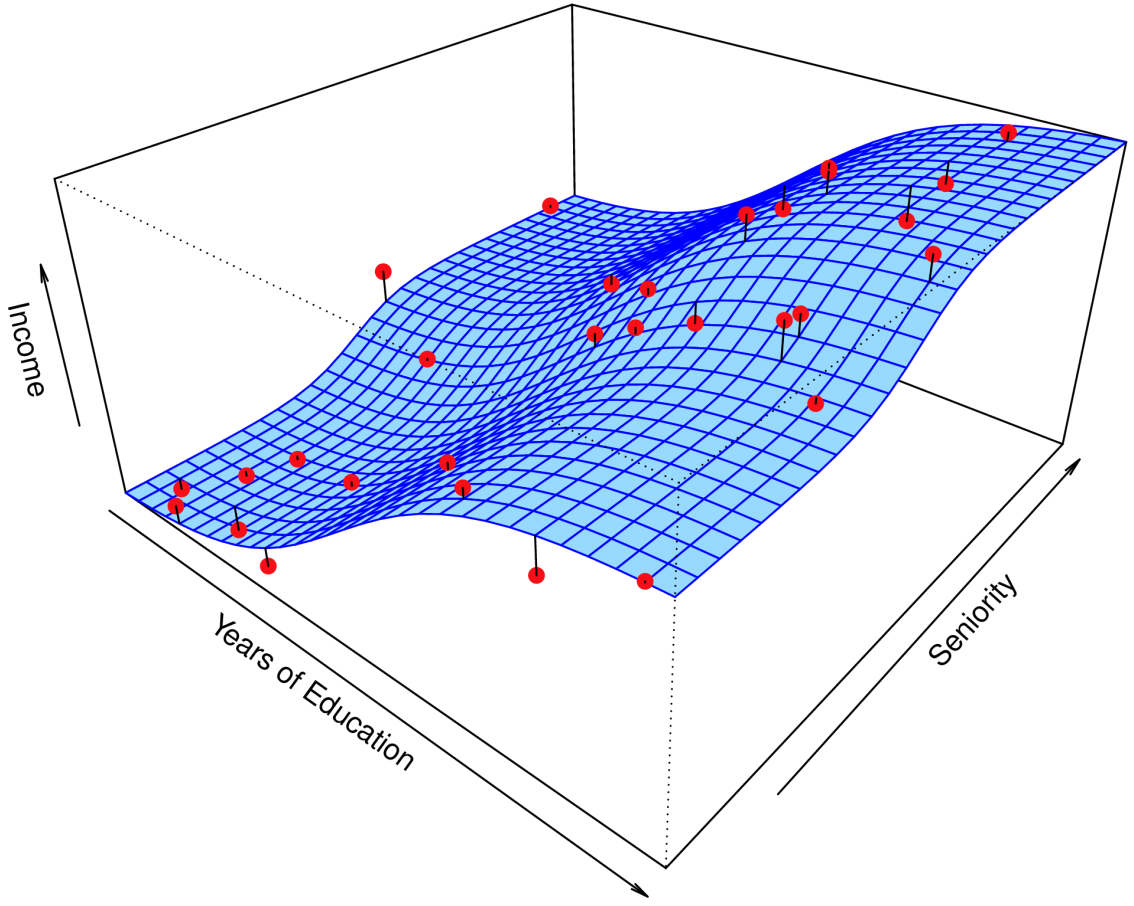
Data in the model space

Wickham, Cook, Hofmann (2015) Visualizing statistical models: Removing the blindfold, Statistical Analysis and Data Mining, 8(4):203-225.



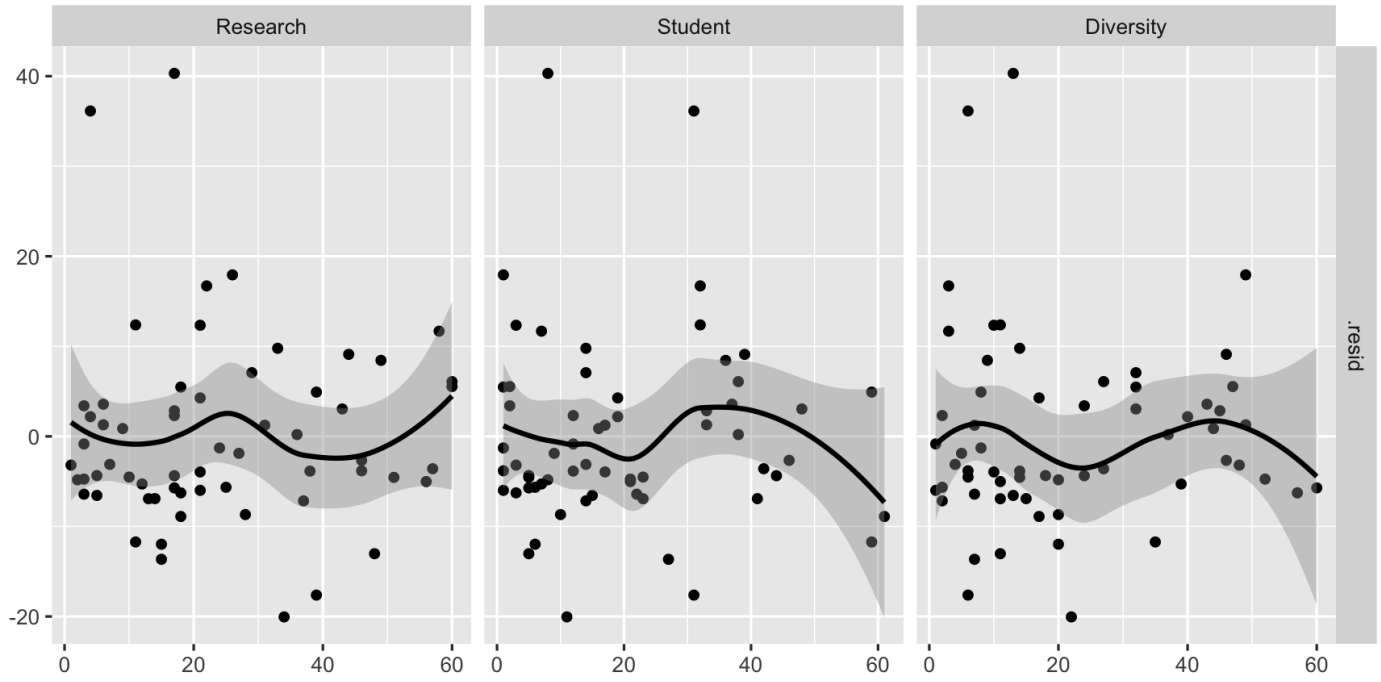
(Chapter3/3.9.pdf)

Model in the data space



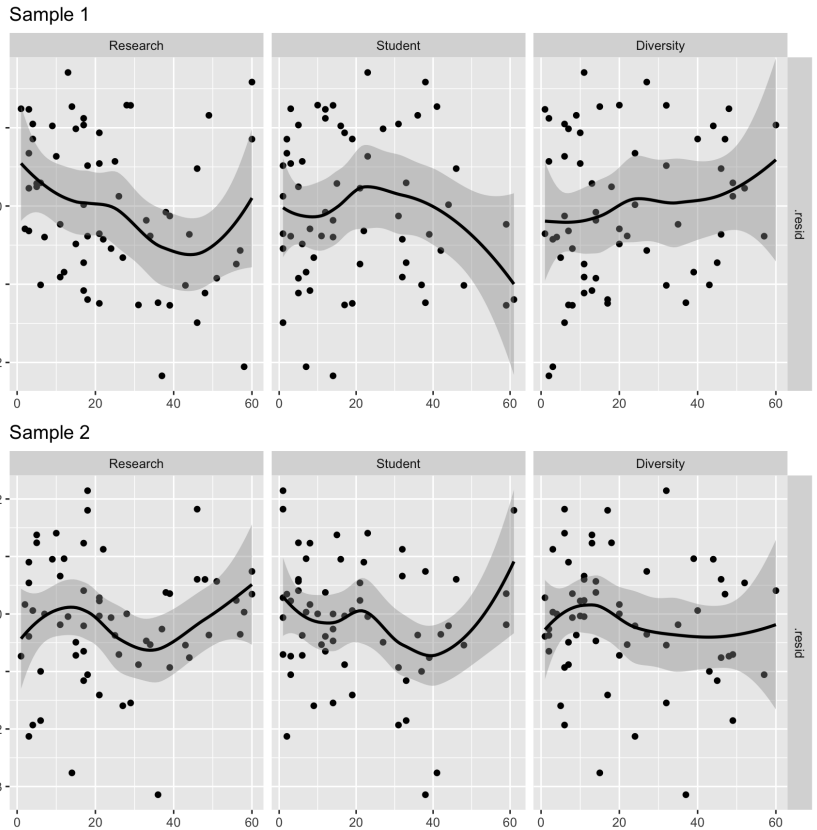
Inference

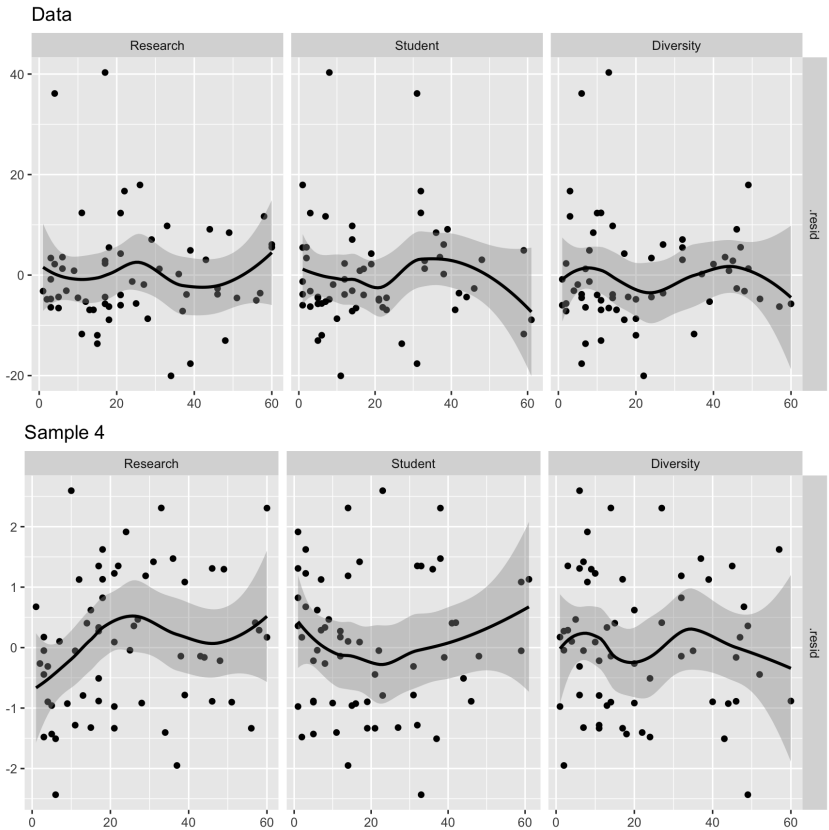
Its hard to read residual plots. How do you know if the residual plot "really" has no structure?



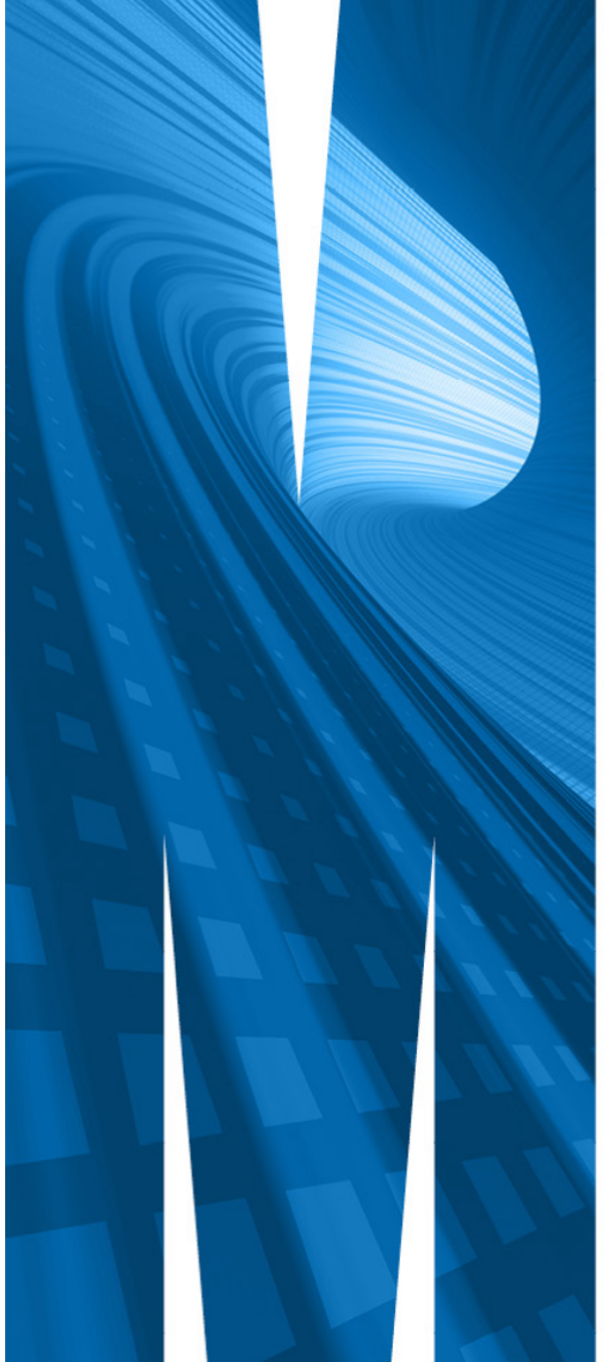
Residual plots with no structure

If there is still structure in the residual plot, the "true" residual plot should be identifiable from the plots of "null" residual plots.





MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 5b