Advanced GDA and Software: Multivariate approaches, Interactive Graphics, Mondrian, iplots and R

Antony Unwin
University of Augsburg
unwin@math.uni-augsburg.de

PolBeRG/ELECDEM Workshop Budapest 28th April, 2012

Why visualize data?

- Looking for global trends
  - overall structure
- Looking for local features
  - data quality
  - groups or clusters
  - outliers, tail distributions and extremes
  - patterns of all kinds

Possible examples

- German Election 2005 (results + demographics)
- German Reichstag Election 1930
- Irish Presidential Election 1990 (last opinion poll)
- Irish Referenda in the 1980s
- "Bowling Alone" US Lifestyle survey over 20 years
- Movies (120,404 films rated on imdb.com)
- Oscar nominees 1928-2000 (age, gender, ...)
- Shipman's victims (UK doctor murdered patients)

German Bundestagswahl 2005

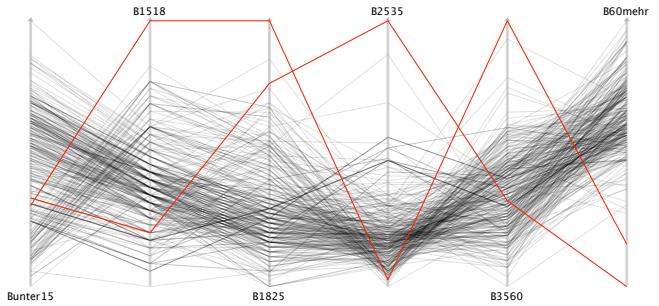
- Votes and % party support for 299 constituencies
  - CDU/CSU, SPD, FDP, Grüne, Linke, Rest
- Accompanying polygon map of the constituencies
- Population demographics
  - gender, age, housing, education, employment …
  - www.bundeswahlleiter.de/de/bundestagswahlen/BTW_BUND_05/strukturdaten/

Germany 2005 questions

- Where are the different parties strongest?
- What associations are there between the parties?
- Are age distributions and unemployment figures associated with party strength?
- Which constituencies stand out locally as being different from their neighbours?

Parallel coordinate plots

- Each variable has its own vertical axis.
- Each case is represented by line segments joining its points on the axes.



PCP options

- Choice of variables
- Scaling and order of the axes (affect the display a lot)
- Rescale axes: inversion, common scaling
- Display as boxplots
- Reorder variables
  - by hand
  - sorting by statistics (max, median, IQ-range ...)
- Interactive tools are important

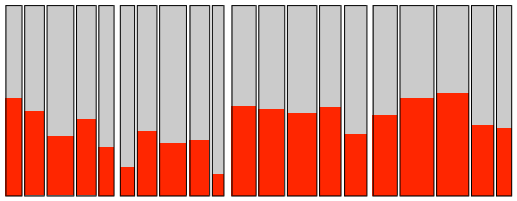
Irish Presidential Election 1990

- Three candidates:
  - Lenihan (FF Foreign Minister, 1982 Phone Scandal)
  - Currie (FG Opposition, from Northern Ireland)
  - Robinson (Labour, first female candidate)
- MRBI Opinion Poll just before the election
- 1000 people asked
  - demographics
  - preferences and influences

Ireland 1990 questions

- Was the survey balanced by sex, age and social class?
- Which groups were strongly for Mary Robinson?
- How influential was party affiliation?
- How crucial for Mary Robinson were the second preferences she got from Austin Currie supporters?
- Four factors were rumoured to be important in determining people's votes. Were they?

Multiple barchart for Area by Sex by Age



Doubledecker plot: Rural/Urban by Sex by Age for Robinson

Mosaicplots: structure

- Variable category combinations are represented by rectangles
- Rectangle area is (usually) proportional to frequency
- Rectangles may have equal width (height), so that height (width) is proportional to frequency
- Rectangles may be aligned in various ways
- Variables may be rotated
- Rectangles may be coloured

Mosaicplots: variants

- Classical (Observed)
- Expected (based on a model)
- Fluctuation diagram
- Same binsize
- Multiple barcharts
- Doubledecker
- rmb
- Weighted mosaics

Mosaicplots: options

- Choice of variables (in effect aggregation)
- Type of mosaicplot
- Order of variables
- Whether variables are plotted horizontally or vertically
- Order of categories within variables
- Aggregation of categories within variables
- Formatting: plot size, aspect ratio, spacing between levels
- Interactive tools are important

What is Interactive Graphics?

- Querying
- Selection, highlighting and linking
- Reformating (rescaling, sorting, colouring, resizing)
- Zooming
- Multiple views
- But: check the probabilities and implicit comparisons

Case study: Movies dataset

- Movies data downloaded from the web (imdb.com)
- Just over 120,000 films
- Information on
  - Year and Length
  - Type (23 binary variables)
  - Average rating and number of votes

Movie questions

- What is the distribution of ratings?
- Do modern films get more votes and higher ratings?
- What sort of ratings do action films get?
- Are short films less often rated than non-shorts?
- What kinds of film get high ratings based on few votes?
- What combinations of film types are there?
- Which film titles occur most often and are these films all from different years?

Case study: Oscars

Redelmeier and Singh (2001) studied the mortality of actors and actresses who had been nominated for the Oscars compared to controls.

There are 1670 cases and 15 variables:

7 demographic variables including
Gender, Year of birth, Year of death

8 film career variables including
films, # four star films, First nomination year

Oscars: Questions

- Are there any data quality issues? How many males and females should there be in the dataset?
- What kinds of stars won when they were young?
- What relationships are there between the numbers of nominations and wins and the numbers of films and fourstar films?
- How have the winners changed over time, if at all?

IG Advantages

- Direct querying
- Multivariate information via linking
- Fast, flexible analyses (including sensitivity analyses)
- Running through alternatives quickly
- Experimental reformatting, versatile reordering
- Generate ideas/hypotheses
- Check implications for other variables
- and
  - don't let computing get in the way of thinking

IG Disadvantages

- Not mathematically defined
- Difficult to record the process
- Cannot replicate analyses
- Difficult to save results of analyses
- Can often not test results statistically
- Not presentation graphics quality
- Data dredging: you always find something

Mondrian

- Mondrian for interactive graphical analysis
  - one of the Augsburg Impressionists
  - stats.math.uni-augsburg.de/Mondrian/
  - for Windows, Unix, MacOS
  - by Martin Theus

Mondrian

- Information
  – http://rosuda.org/Mondrian/
- Further help
  – the reference card
- Plots
  – missing value, histogram, boxplot, barchart, scatterplot, splom, mosaicplots, parallel coordinate plot, map
- Rserve is necessary to use R from Mondrian
  – density estimation, CDPlot, smoothers

Mondrian — features

- Querying
- Selection and highlighting (incl. selection sequences)
- Zooming
- Rescaling
- Resizing points
- Sorting
- Colouring
- alpha-blending
- Printing

R and Graphics

- Base graphics v. *grid*
- Packages include (cf. also the Graphics Task View)
  - *vcd* :for displaying categorical data
  - *ggplot2* :implementation of "Grammar of Graphics" including `qplot`
  - *lattice* :for drawing trellis plots
  - *iplots* :for interactive graphics

iPlots and R

- Uses the JGR interface to R
  - http://stats.math.uni-augsburg.de/JGR/
- iPlots is an interactive graphics R package
  - http://stats.math.uni-augsburg.de/iPlots/
- ibar, ihist, iplot, imosaic, ipcp
- Query with the ctrl key
- Options available from the View menu
- Developed primarily by Simon Urbanek

Summary

- Parallel coordinate plots are for multivariate continuous data
- Mosaicplots are for multivariate categorical data
  - both have to be interactive
- Interactive graphics are for exploring data
  - becoming used for web presentations in a limited way
- Graphics require interactive thought!

Case study: Shipman dataset

In 2000 the British doctor, Harold Shipman, was convicted of murdering 15 of his patients. The official report (www.the-shipman-inquiry.org.uk/), which examined the deaths of all patients under his care over twenty years, concluded that he had probably murdered over 200. Details of the deaths of 508 of his patients where there was doubt about the cause of death have been taken from Appendix F of the report.

Variable	Description
ID	patient number
Day	day of death
Month	month of death
Year	year of death
Weekday	day of week of death
Date	days since 1/1/1904
Name	full name of patient
Surname	surname of patient
Sex	gender of patient
Age	age at death
Location	place of death
Decision	official view on Dr.'s guilt

From the Appendix

APPENDIX F **Chronological List of Decided Cases**

Date of Death	Name of Deceased	Age of Deceased	Place of Death	Decision
1974				
10/5/74	Ruth Highley	72	Own home	Natural death
22/6/74	Edith Annie Bill	67	Own home	Natural death
23/7/74	Colin Whitham	26	Own home	Natural death
2/8/74	Stanley Uttley	58	Surgery	Natural death
9/10/74	Hena Cheetham	77	Ambulance	Natural death
10/11/74	Harold Edward Jackman	78	Hospital	Natural death
9/12/74	Sean Stuart Callaghan	18	Hospital	Natural death
16/12/74	Moira Kelly	26	Hospital	Natural death
29/12/74	Sarah Ann Thomas	86	Own home	Insufficient evidence for decision
There is also a decision in respect of Frances Elaine Oswald, relating to an incident which took place on 21/08/74				
1975				
21/1/75	Lily Crossley	73	Own home	Suspicion of unlawful killing
21/1/75	Robert Henry Lingard	62	Own home	Suspicion of unlawful killing

Shipman questions?

- Which variables might be most useful? Draw plots to look for patterns.
- Were there periods when there were there no deaths/murders? Was there a pattern in the deaths by day of the week?
- Is there any pattern in the age and gender of the victims?
- Was the place of death relevant?