Lecture 3 + Seminar 3

A workshop on graphing using ggplot2

Plotting packages in R

- Graphics functions in the `base` package: `plot`( ), `hist`( ), etc.
- `lattice` package
- `ggplot2`

Have an internal logic for a range of statistical plots

- Specialised functions inside other packages: `plotrix`, `car`, `plotmeans`, etc.

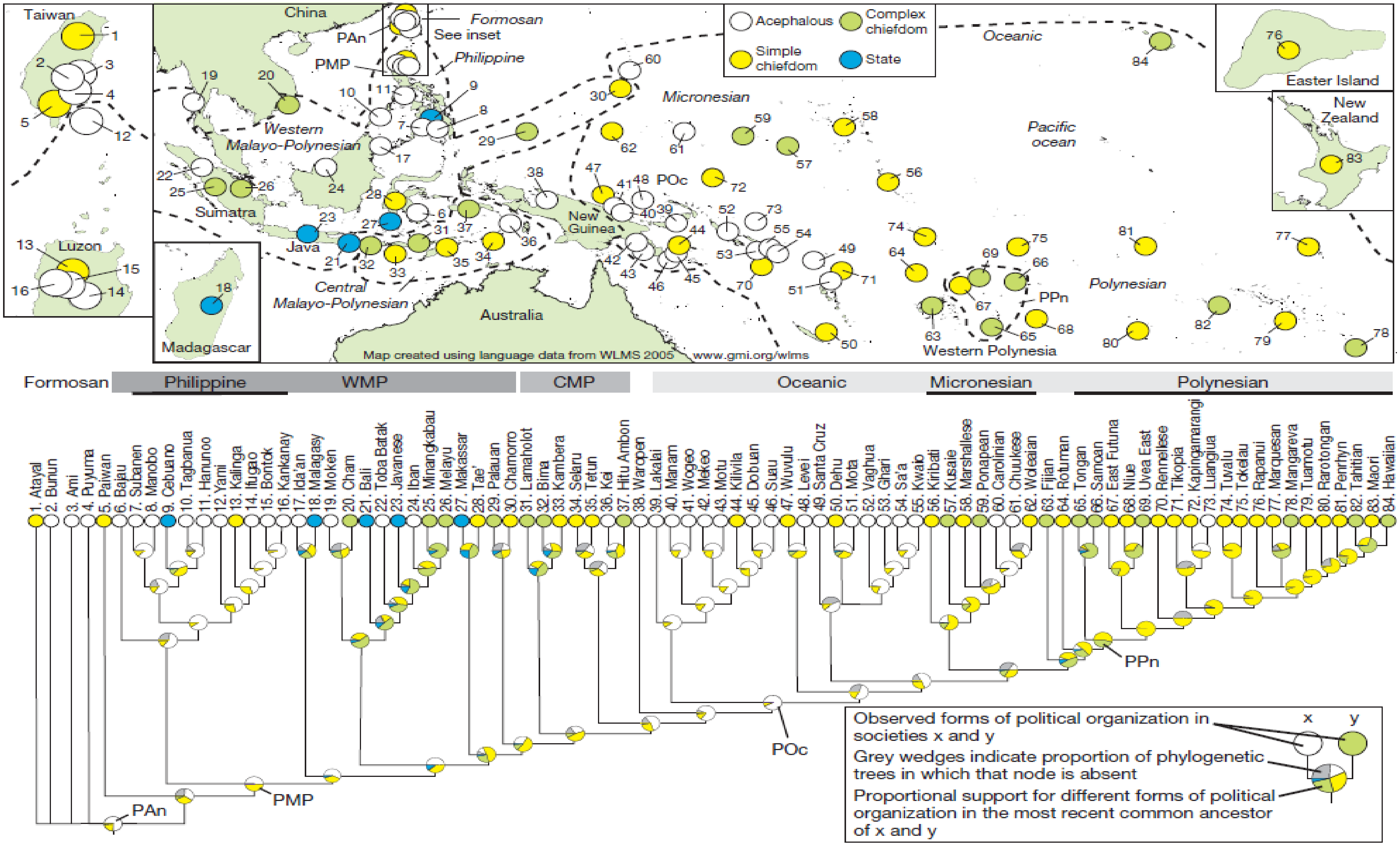
Programme

- Graphing and figures – general principles
- Guidelines for different kinds of graphs and publications:
  - Specialist (colour) publications
  - Presentations
  - Journal papers (APA guidelines)
- Quick plots using `qplot` (plus, a review on saving graphs)
- More flexible graphs using `ggplot`
- Final formatting using `ggplot`
- Seminar and assignment: At the seminar, you will complete an exercise to create an APA-style plot. Then, to complete the assignment, you will need to add at least one additional piece of formatting at home.
- Reading: book by the author of `ggplot`

General principles: Graphs and figures should…

1. Summarise and/or reveal data, making large datasets coherent
2. Encourage the viewer to think about the data being presented (rather than some aspect of the graph, like how pink, “pretty” or poorly visible it is)
3. Avoid distorting the data
4. Encourage the viewer to compare different pieces of data

A graph that satisfies all four criteria despite its complexity:



Observed forms of political organization in societies x and y

Grey wedges indicate proportion of phylogenetic trees in which that node is absent

Proportional support for different forms of political organization in the most recent common ancestor of x and y

Figure 1 | Phylogenetic relationships and geographical location of 84 Austronesian societies. Majority-rule consensus tree summarizing the 1,000 Austronesian basic vocabulary trees (major linguistic sub-groupings indicated) (see also Supplementary Fig. 1). Circles at the nodes in the consensus tree indicate proportional levels of support for different forms of political organization in ancestral societies under the best-fitting model of evolution (unilinear). Nodes relating to often discussed ancestral Austronesian communities are highlighted: proto-Austronesian (PAn; ancestor to all Austronesian societies), proto-Malayo-Polynesian (PMP), proto-Oceanic (POc) and proto-Polynesian (PPn). Approximate locations of these ancestral communities are indicated on the map. Branch lengths in this figure are arbitrary (that is, not proportional to the degree of linguistic change or time). WMP, Western Malayo-Polynesian; CMP, Central Malayo-Polynesian.

A graph that satisfies all four criteria despite its simplicity:

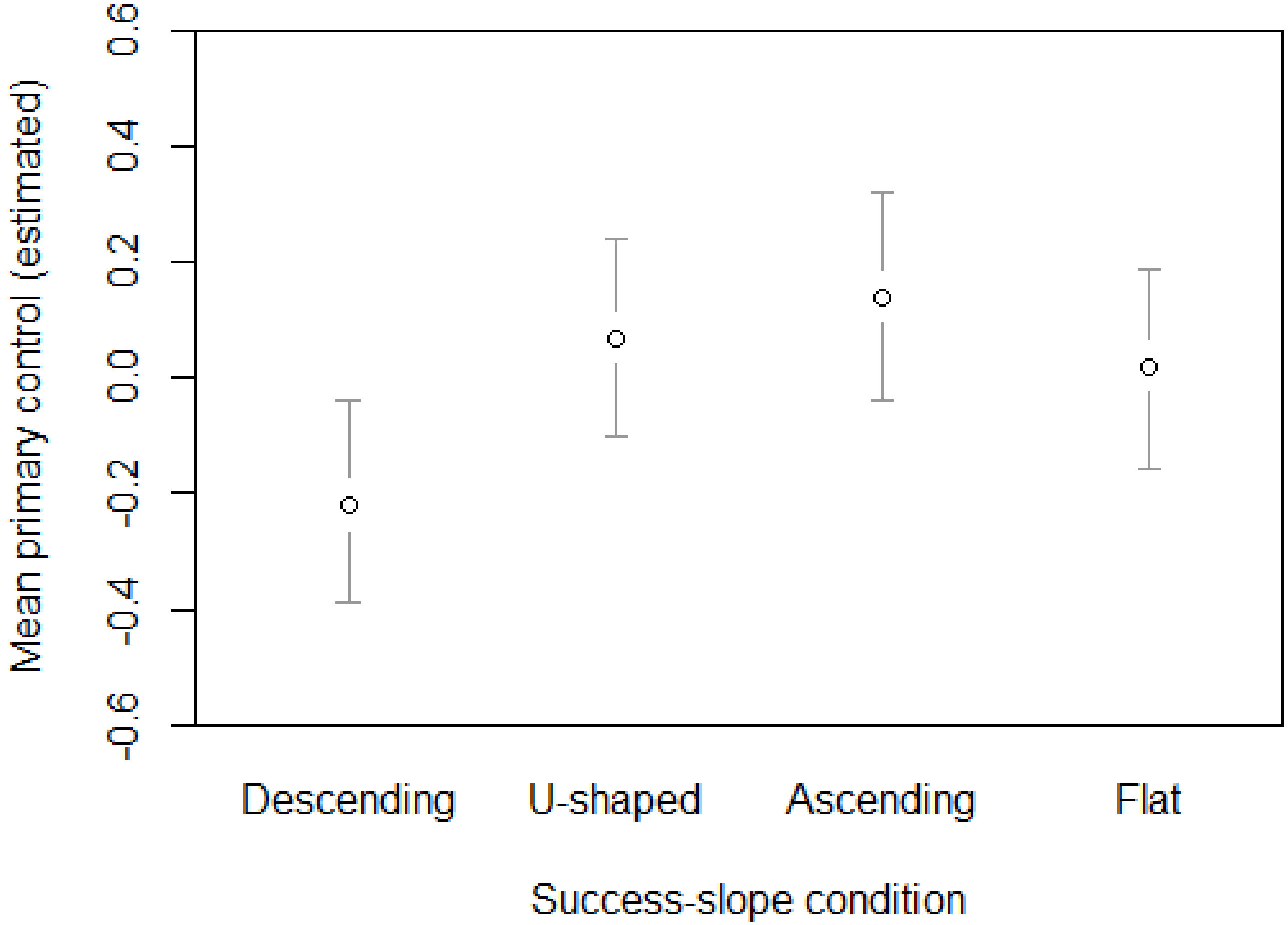


Figure 4.2. Estimated means (and 95% CIs) of inferred natural control across success-slope conditions when covariates (soccer interest and prior beliefs) were evaluated at their mean values.

Guidelines for different types of graphs

- All graphs:
  - Think about when to start scales at the origin (0)
- Bar graphs:
  - If needing colour, use pastel colours to avoid overwhelming the viewer.
  - Best to use some shading (e.g., grey or patterns) to prevent the graph from looking like a wire frame.
  - Histograms: consider displaying the mean/median value as well.
- Graphs involving points (and possibly lines):
  - Vibrant colours work well for points and lines, making them quite visible.
  - Best to start y-axis at 0.
  - Use a line if there is a flow of time from one point to the next.
  - Patterns disappear if you stretch the x-axis too far.
- Scatter plots:
  - Simplify the look by not using a thick border.
  - Any line through the dots must be thicker than the dots or in a brighter colour.

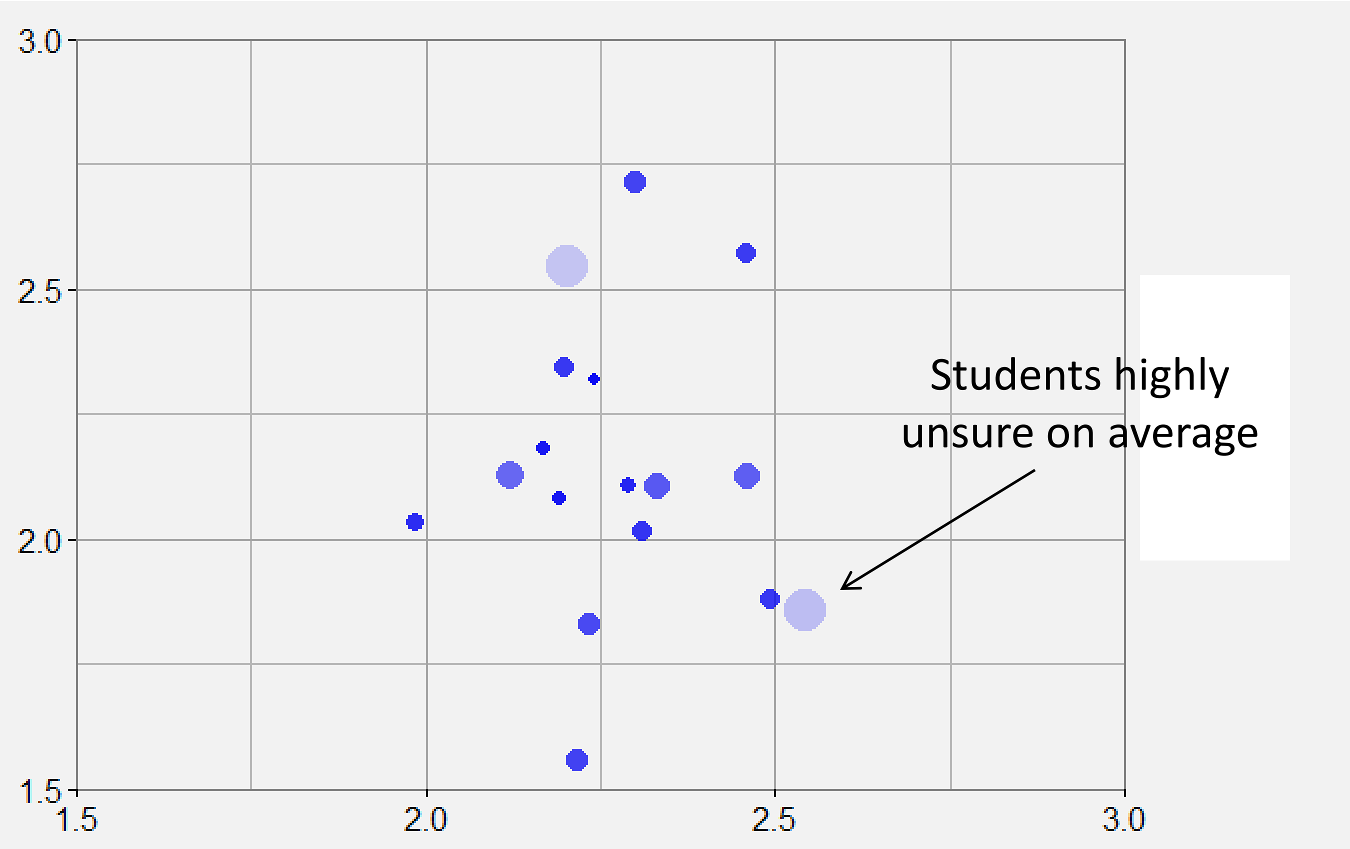
Guidelines for different publication types

Think about:

- What story do you want to tell?
- Who are you telling the story to?

Journal articles	Specialist (colour) articles	Conference presentations
Follow APA guidelines – e.g., page 4 in this style manual.	• Follow APA guidelines as much as possible, but use colour if an additional explanatory tool is needed. • Colours: think about whether they should be vibrant or neutral, given (a) your topic, and (b) whether you are plotting points or bars (see previous slide).	• No caption available, so more labelling in the title or inside the graph is needed. • Figures can be useful for a non-specialist audience. • Same advice as for specialist articles regarding colour. • For small sample sizes, you might not even need an x-axis and y-axis. Can instead label points of interest directly.

Highlighting individual data points:



Students highly unsure on average

University ratings by foreign students

University ratings by Czech students

Quick plots using `qplot`

- The `qplot` function is in the `ggplot2` package.
- The function is very useful for data exploration, as it is possible to draw fairly complex plots with one or two lines of code.
- The function is not useful for final plots for presentations and publications because the overall appearance of the plots is difficult to change.
- Basic principle: "geoms" (representations of data) have "aesthetics“ (properties) that can be "mapped" to variables in the dataset or “set” to a desired value
  - geom examples: point, histogram, smooth (regression line)
  - aesthetics (aes) examples: x and y (the variables being plotted), colour, size, shape, alpha (transparency), group
  - additional arguments to do with avoiding overplotting: position and facet (facet_grid and facet_wrap)

Start script

Saving graphs: a review

- In the Plots tab, Export -> Save Plot As Image... Then choose Image Format and Size
- By default, graphs are saved to your working directory, but you can choose any folder by clicking “Directory” after clicking “Save Plot As Image”.

Back to script

More flexible graphs using `ggplot`

- The `ggplot` function is also in the `ggplot2` package.
- Key concepts, apart from the already mentioned geoms, aesthetics, position, facet, setting and mapping:
  - Layers (+): The graph is not displayed until you add a layer, but it is customary to specify the aesthetics that apply to all layers at the very beginning.
  - Stats: Stats have default geoms, while geoms have default stats.
  - Each plot is treated as a variable (e.g., sepf3).
  - The aesthetics in each layer override any aesthetics specified at the beginning.
  - Search for these terms in the script and in the book for concrete examples.
- We covered: overlaying of histograms and regression lines, adding error bars to line plots and bar plots, setting axis limits (coord_cartesian), faceting, and free scales.

Final formatting using ggplot

	Adjusted through	**Terms to find in the script**

Overall colour-scheme:
• black and white?
• settings for colours
	theme_set
	scale_colour_manual
		theme_bw()
		scale_colour_hue()
		scale_colour_grey

Appearance of points	geom_point geom_params	
Appearance of error bars	stat_summary (in our script) Other possibilities: geom_errorbar, geom_params	colour = "gray41"
Appearance of lines	scale_linetype_manual	scale_linetype_manual(values=c("dotted", "solid", "longdash", "dotdash"))

Gridlines	theme(panel.grid.major = element_line( ))
	theme(panel.grid.minor = element_line( ))
		panel.grid.major = element_line(colour = "gray41", size = 1)
		panel.grid.minor.y = element_blank()

	Adjusted through	Terms to find in the script

Labels along the axes	scale_x_continuous (when x is not a factor variable)
	scale_x_discrete
	scale_y_continous
	scale_y_discrete
		scale_x_continuous(breaks = 1:2, labels = c("Trials 1-24", "Trials 25-48“, name = "Time period")
		scale_y_continuous(name = "")

Facet labels	theme(strip.text = ___)
	Changes to name of factor levels
		strip.text.y = element_text(size=14, face = "bold")
		levels(longsub2$Measure) <- c("Kick Dir Entropy", "No. of Player\nChanges")

Text size	theme(___ = element_text( ))	theme(axis.title.x = element_text(size= 20), axis.text.y = element_text(size=14, colour = "black")

Legend	scale_linetype_manual
	scale_colour_hue
	etc. Depending on what aesthetic (colour, linetype, shape) you have mapped the variable to
	theme(legend.text = ___)
		scale_linetype_manual(values=c("dotted", "solid", "longdash", "dotdash"), name="Success Slope", breaks=c("Descending", "U-shaped", "Ascending", "Flat"), labels=c("Desc.", "U-shaped", "Ascending", "Flat"))
		legend.text = element_text(size=12)

Layering during final formatting (Script line 597)

If you are adding something to an existing layer – such as stat_summary( ) – you need to redraw the whole plot under its current name or a new name.

```
sepf3 <- ggplot(data = longsub2, aes(x=Time, y=Value, group = SeqCond)) +
   geom_line(mapping = aes(linetype = SeqCond), stat = "summary",
           stat_params = list(fun.y = mean), position = pd,
           geom_params = list(size = .75)) +
   facet_grid(Measure ~ ., scales = "free_y") +
   stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
             ymin = min,
             ymax = max,
             width = .05,
             mult = 1,
             position = pd,
             colour = "gray41") +
   geom_point(stat = "summary", stat_params = list(fun.y = mean), position = pd,
            geom_params = list(fill = "white", shape = 21, size = 3)) +
   scale_x_continuous(breaks = 1:2, labels = c("Trials 1-24", "Trials 25-48"),
                  name = "Time period") +
   scale_linetype_manual(values=c("dotted", "solid", "longdash", "dotdash"))

sepf3
```

Layering during final formatting

If you are adding an entirely new layer, you can just use “+”, but it is best to avoid this, since you are highly likely to need to go back and make many changes to existing layers later on.

```
sepf4 <- sepf3 + theme(axis.title.x = element_text(size= 20))
```

```
sepf5  <- sepf4 + theme(axis.text.y = element_text(size=14, colour = "black"),
        axis.text.x = element_text(size = 14, colour = "black"),
        strip.text.y = element_text(size=14, face = "bold"))
```

There would be an error drawing this because some aspects of theme ( ) were already specified when drawing sepf4.

Reading

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Available online: http://moderngraphics11.pbworks.com/f/ggplot2-Book09hWickham.pdf

Additional resources:

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R.* Sage: UK. Chapter 4. Exploring data with graphs.

Support website: http://docs.ggplot2.org/current/