DATA VISUALISATION

UNIVERSITÄT
LEIPZIG

Erik Kusch

erik.kusch@i-solution.de

Section for Ecoinformatics & Biodiversity
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)
Aarhus University

Means to Visualisation

Methods of data visualisation are manifold:

Tables:

- Data Tables
- Frequency Tables
- Stem And Leaf Displays

Text-based descriptions of data:

- Only applicable to minute data sets
- Not used extensively

Plots:

- Pie Charts
- Scatter plots, Line Graphs
- Bar Charts, Histograms, Frequency Polygons
- Box plots
- Contour Plots, 3-D Plots
- ...

$\rightarrow$ We will not be covering text-based data summaries here.

Table Etiquette

Tables are **useful data summary and visualisation tools**.

Etiquette in table making:

- Vertical lines are used sparingly
- Horizontal lines are used frequently
- Table captions are placed **above** the table they belong to

Making tables directly in R can be difficult. Assuming you use LaTeX for writing manuscripts (which you really should try if you haven't yet):

- LaTeX directly
- The Excel2LaTeX-Add-in (https://www.ctan.org/tex-archive/support/excel2latex/)
- Various R packages (e.g.: 'xtable')
- R Markdown for writing manuscripts

Data Tables

Can accommodate all kinds of data.

- **For publications:**
  - Great way to summarise and present data
  - Can be used to present a list of definitions

Table 1 Selected definitions of resilience that have been proposed in the ecological literature

Definition	Source
The magnitude of disturbance that can be tolerated before a system moves into a different region of state space and a different set of controls	Carpenter *et al.* (2001)
The ability of the system to maintain its identity in the face of internal change and external shocks and disturbances	Cumming *et al.* (2005)
The capacity of a system to absorb disturbance and reorganize while undergoing change so as to still retain essentially the same function, structure and feedbacks, and therefore identity, that is, the capacity to change in order to maintain the same identity.	Folke *et al.* (2010)
Returning to the reference state (or dynamic) after a temporary disturbance	Grimm & Wissel (1997)
Resilience refers to the width or limit of a stability domain and is defined by the magnitude of disturbance that a system can absorb before it changes stable states	Gunderson (2000)
Resilience determines the persistence of relationships within a system and is a measure of the ability of these systems to absorb changes of state variables, driving variables, and parameters, and still persist	Holling (1973)
How fast the variables return toward their equilibrium following a perturbation.	Pimm (1984)
The ability of the system to return to the original state after a disturbance.	Scheffer *et al.* (2002)
Helpful resilience: Resilience that helps to maintain a predisturbance ecosystem state so that it does not cross a threshold.	Standish *et al.* (2014)
Unhelpful resilience: Resilience that helps to maintain an ecosystem in a degraded state following a disturbance.	
The capacity of a system to experience shocks while retaining essentially the same function, structure, feedbacks, and therefore identity	Walker *et al.* (2006)
The capacity of a system to absorb disturbance and reorganize while undergoing change so as to still retain essentially the same function, structure, identity, and feedbacks.	Walker *et al.* (2004)

Newton, A. C. (2016) 'Biodiversity Risks of Adopting Resilience as a Policy Goal', Conservation Letters, 9(October), pp. 369-376. doi: 10.1111/conl.12227.

Data Tables

Can accommodate all kinds of data.

- **For publications:**
  - Great way to summarise and present data
  - Can be used to present a list of definitions
- **For behind-the-scenes work:**
  - Still a great way to summarise and present data
  - Data management, mining and exploration relies on tables (more on this in seminar 7)

Table 1 Comparison of 61 present-day temperatures and Middle Pliocene temperature estimates from the general circulation model (GCM) and literature for selected regions (see Supplementary Appendix S1 for references). *t*-tests assuming unequal variance are applied for the comparison of model and Middle Pliocene temperature reconstructions with $n > 2$ (n.s. = not significant; n.a. = not applicable due to small sample size). Key to units: MAT, mean annual temperature; MAWT/ST, mean annual winter/summer temperature.

				Temperature (°C)			
Region	Site	*n*	Unit	Present	Model	Palaeodata	Significance
Alaska, E Siberia	3, 5–7, 135, 136	6	MAT	−9.1 ± 4.1	−3.5 ± 5.0	2.3 ± 0.6	$P < 0.05$
East Siberia	135–136	2	MAWT	−30.3 ± 6.8	−15.0 ± 0	−15.4 ± 14.8	n.a.
		2	MAST	6.7 ± 4.7	15.5 ± 0	6.0 ± 8.3	n.a.
West Siberia	132	1	MAT	−0.7	7.8	16.0	n.a.
Labrador/Quebec	14	1	MAT	2.2	6.0	5.7	n.a.
Iceland	50, 51	2	MAT	0.8 ± 1.2	2.1 ± 0.3	3.7 ± 0.4	n.a.
Russian Plain	59–62, 66	5	MAWT	−13.1 ± 2.8	−3.6 ± 3.2	−1.1 ± 0.8	n.s.
		5	MAST	16.9 ± 1.6	22.9 ± 2.7	19.6 ± 0.5	$P < 0.05$
Germany	70, 72, 74	3	MAT	8.0 ± 0.1	13.8 ± 0.4	11.5 ± 2.8	n.s.
Black Sea	84, 86, 89	3	MAWT	−3.7 ± 5.2	7.0 ± 3.6	0.7 ± 4.2	n.s
		3	MAST	21.3 ± 0.8	26.0 ± 5.8	21.7 ± 1.0	n.s.
Azerbaijan	93	1	MAT	12.1	18.0	17.0	n.a.
West coast USA	17, 25	2	MAT	10.5 ± 4.5	14.3 ± 1.3	15.5 ± 3.5	n.a.
East USA	33	1	MAT	14.9	18.8	17.2	n.a.
N Mediterranean	79, 94, 98–99	4	MAT	13.3 ± 3.3	16.6 ± 2.5	17.5 ± 1.5	n.s.
S Mediterranean	95, 97, 108, 109	4	MAT	15.4 ± 1.2	18.3 ± 2.0	22.2 ± 1.0	$P < 0.05$
Central America	24, 25, 39–40	4	MAT	25.0 ± 1.6	25.5 ± 1.7	25.2 ± 1.8	n.s.
East Africa	118	1	MAT	25.6	20.3	20.2	n.a.
China, Shanxi	157	1	MAT	7.0	12.5	5.0	n.a.
China, Yunnan	140	3	MAT	14.8 ± 0.8	14.0	17.8 ± 2.3	$P < 0.05$
Japan	163	1	MAT	16.6	21.9	18.0	n.a.
NE Australia	183, 184	2	MAT	24.1	26.2	20.0 ± 2.8	n.a.
SE Australia	187, 190, 193	3	MAT	13.9 ± 2.4	15.2 ± 2.6	15.3 ± 2.8	n.s.
Antarctica	201	1	MAT	−47.0	−33.0	−12.0	n.a.

Salzmann, U. et al. (2008) 'A new global biome reconstruction and data-model comparison for the Middle Pliocene', Global Ecology and Biogeography, 17(3), pp. 432-447. doi: 10.1111/j.1466-8238.2008.00381.x.

Frequency Tables

Only accommodate frequency counts.

- **For publications:**
  - Rarely ever used in publications
  - Applicable for publication of appendices and manuscripts of theses
- **For behind-the-scenes work:**
  - Used excessively internally in ‘R‘
  - Basis for many plotting approaches

Frequency of IGBP classes in the EDC, UMD and BU maps

	EDC		UMD		BU	
Class	Pixels	%	Pixels	%	Pixels	%
1	3736947	17.0	2306360	10.5	3472296	15.9
2	354661	1.6	338431	1.5	356568	1.6
4	1488111	6.8	778641	3.6	1574119	7.2
5	2854132	13.0	1196545	5.5	948452	4.3
6	579582	2.6	1656401	7.6	398707	1.8
7	2306720	10.5	2970955	13.6	5047635	23.0
8	1571191	7.2	3263042	14.9	1469668	6.7
9	73694	0.3	3111528	14.2	316218	1.4
10	1658740	7.6	1977154	9.0	963878	4.4
11	359708	1.6				
12	1852240	8.5	1818480	8.3	3693475	16.9
13	84539	0.4	84539	0.4	84539	0.4
14	1510139	6.9			1497631	6.8
15	1472376	6.7			1555290	7.1
16	1998884	9.2	2399588	11.0	605572	2.8
Total land mass = 21899509						

Table 20: Frequency of classes in the IGBP scheme for the UMD, EDC and BU maps.

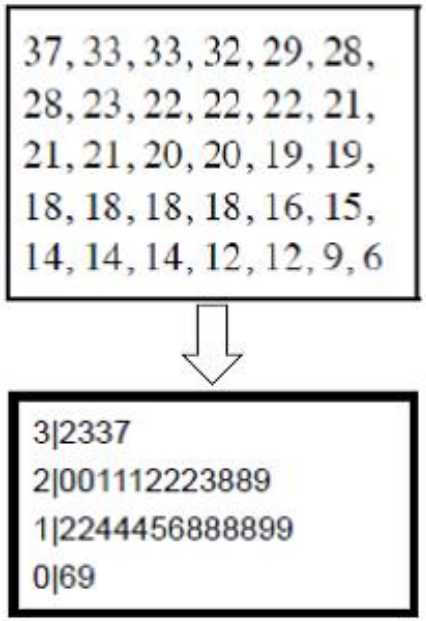
Lotsch, A. (1996) Biome level classification of land cover at continental scales using decision trees. Free University of Berlin. Available at: http://cliveg.bu.edu/download/thdis/alotsch.MA.pdf.

Note that the table caption is misplaced on this table.

Stems And Leaf Displays

Accommodate frequency/count data.

- **For publications:**
  - Pretty outdated
  - Usually only included in books and course material
- **For behind-the-scenes work:**
  - Of no particular use when considering small or excessively big data sets
  - Can be helpful in data exploration of medium-sized data sets



Lane, D. M. (2009) Introduction To Statistics, Introduction to Statistics. doi: 10.1016/B978-0-12-370483-2.00006-0.

Plot Etiquette

Plots are extremely **useful data summary and visualisation tools**.

Etiquette in plot making:

- Less is more (strife for simplicity)
- Figure captions are placed **below** the figure they belong to

Making plots directly in `R` entails a learning curve and there is a heavy debate about how to do the plotting:

Using **base R**:

- Can be cumbersome
- Relies on same commands as basic R coding

Using **ggplot**:

- Extremely powerful
- Relies on ggplot specific commands

The Good, The Bad, And The Ugly

- **Good plots:**
  - Clearly legible labels
  - Clean look
  - Concise caption
- **Bad plots:**
  - Convoluted display
  - Overlapping plotting symbols
  - Overly complicated caption
- **Ugly plots:**
  - Photos
    - No-Go for publications
    - Ok for presentations
    - Very good for keeping track of complex set-ups for yourself and to aid memory when doing field work
  - Awkward legend
  - Awkward labelling (e.g. obvious R internal naming)
  - Excel figures

The Good

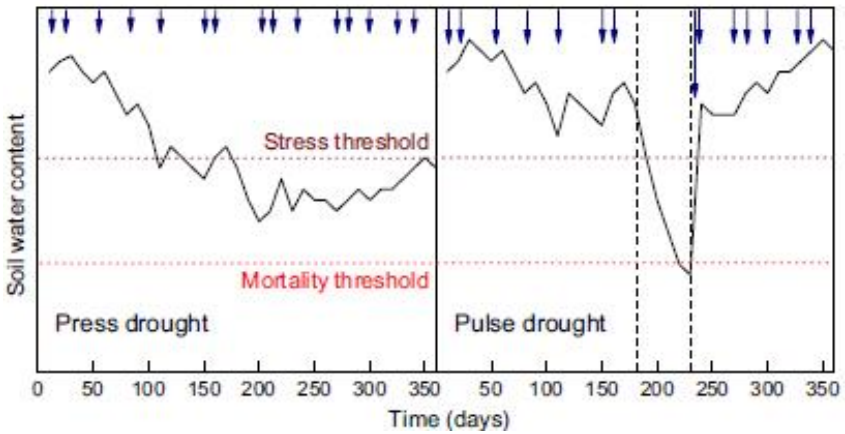


FIGURE 2 Conceptual depiction of a press drought (extended period with sparse precipitation) and a pulse drought (short period with no significant precipitation), identical in return time (extremity). Stress (hypothetical thresholds for species × indicated) reaches less extreme levels during press droughts, but lasts longer and features only short periods when (limited) recovery is possible. Precipitation events are depicted by arrows with a length that scales with precipitation amount

De Boeck, H. J. et al. (2017) 'Patterns and drivers of biodiversity-stability relationships under climate extremes', Journal of Ecology, (October), pp. 1-13. doi: 10.1111/1365-2745.12897.

The Bad

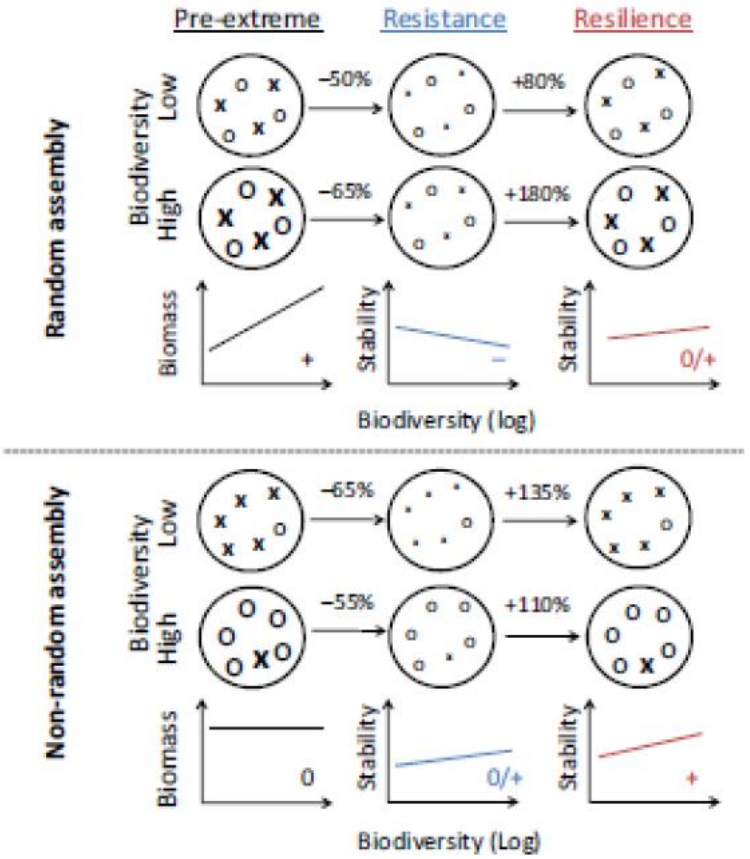


FIGURE 4 A hypothetical example illustrating potential interactions between biodiversity–stability relationships and patterns of community assembly. Two types of species (X and O) with contrasting levels of productivity are present: X-type species are twice as productive as O-type species under normal conditions. Biomass production, relative to each species' type inherent production, is represented by font size, and relative changes in community biomass are indicated. Random assembly implies that both species types are equally represented at all diversity levels, while non-random assembly results in a higher proportion of productive species in low-diversity communities (which would be expected under nutrient enrichment, see text). Less productive species are assumed to be more resistant, but to have slow recovery. The effects of biodiversity on stability during and after the extreme events are shown, assuming a limited recovery period. Resilience integrates both resistance and recovery. In this example, positive effects of biodiversity (pre-extreme includes overyielding) are assumed to decrease during the resistance phase, and increase during the recovery phase (cf. DeClerck et al., 2006; Van Ruijven & Berendse, 2010). Details of the calculations can be found in Table S3. Figure S2 depicts results without different diversity effects during resistance and recovery

De Boeck, H. J. et al. (2017) 'Patterns and drivers of biodiversity-stability relationships under climate extremes', Journal of Ecology, (October), pp. 1-13. doi: 10.1111/1365-2745.12897.

The Ugly

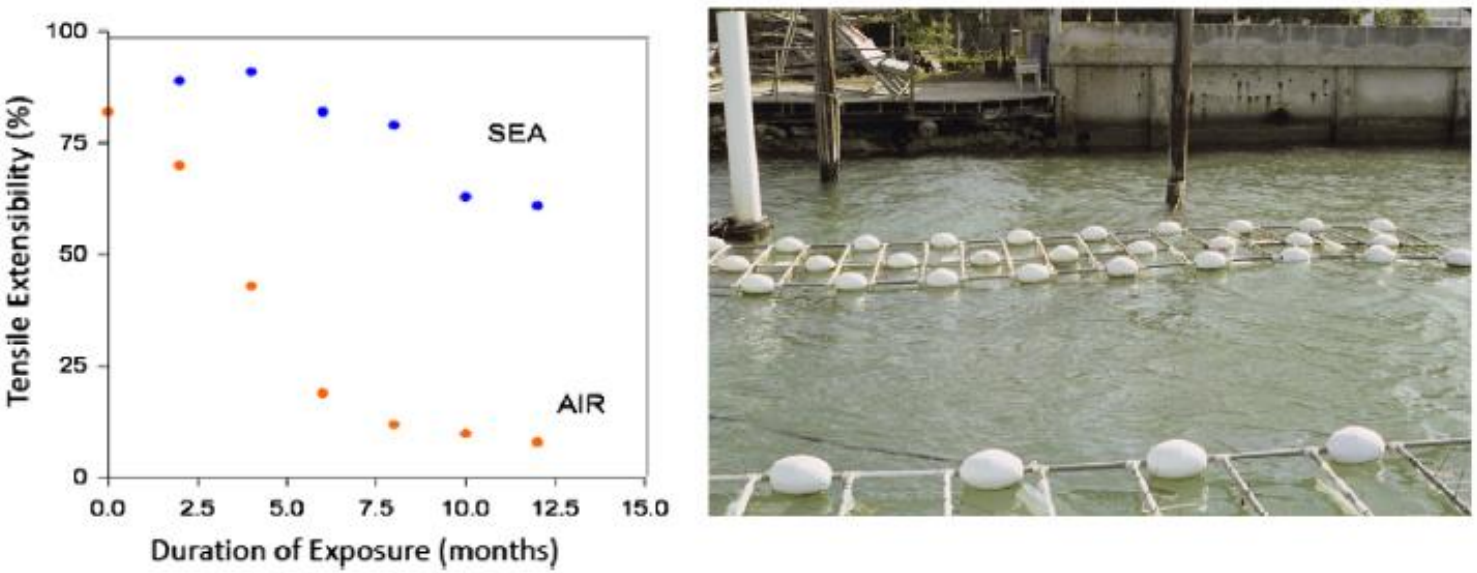


Fig. 2. (Left): Change in percent original tensile extensibility of polypropylene tape exposed in air and floating in sea water in Biscayne Bay, FL. (Right): The floating rig used to expose plastics to surface water environment (Miami Beach, FL).

Andrady, A. L. (2011) 'Microplastics in the marine environment', Marine Pollution Bulletin. Pergamon, 62(8), pp. 1596-1605. doi: 10.1016/J.MARPOLBUL.2011.05.030.

The Ugly

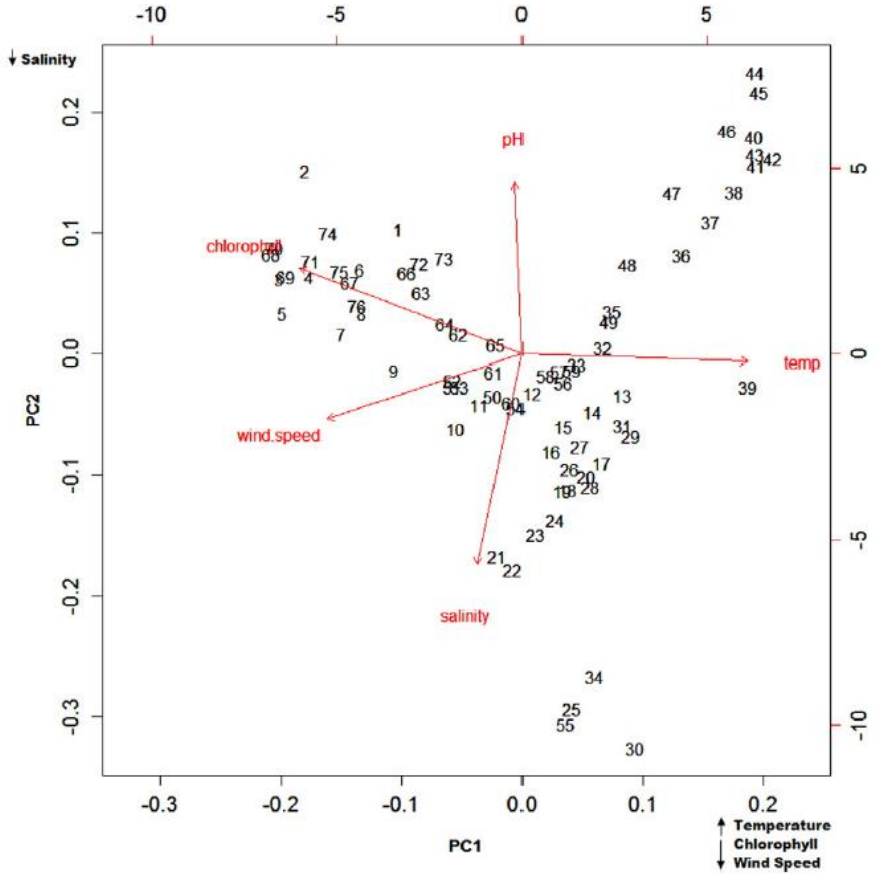


Fig. 1. Biplot showing sampling sites based on environmental variables.

Kanhai, L. D. K. et al. (2017) 'Microplastic abundance, distribution and composition along a latitudinal gradient in the Atlantic Ocean', Marine Pollution Bulletin, 115(1-2), pp. 307-314. doi: 10.1016/j.marpolbul.2016.12.025.

ggplot Overview

These seminars will focus on how to create plots with ‘ggplot‘ instead of teaching you data visualisation using base ‘R‘ commands.

Why we use ggplot:

- It is extremely powerful
- It is becoming the norm
- Even base graphics look good

Why ggplot can frustrate you:

- You need to memorise specific commands
- Certain objects in ‘R‘ are not compatible with ‘ggplot‘ yet
- It may be unintuituve at first

If you need an introduction to base plot, go here:
https://biostats.w.uib.no/topics/r/r-7-making-plot-learn-the-basics/.

How To Make A Plot In R (using ggplot)

The `ggplot()` function considers **three basic components** to a plot:

- *Data set* - where to get the data to the plotted from

```
ggplot(diamonds)
```

- *Aesthetics* - what variables should be used

```
ggplot(diamonds, aes(x=carat, y=price))
```

- *Layers/Geometry* - what kind of plot to produce

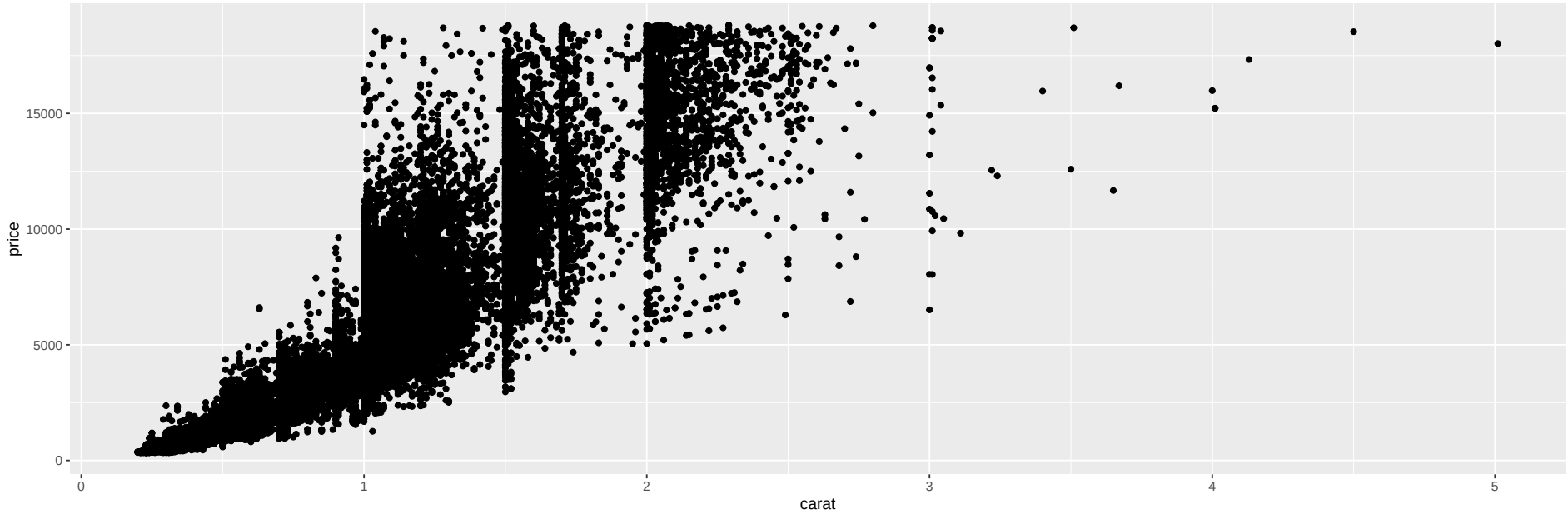
```
ggplot(diamonds,aes(x=carat, y=price)) + geom_point()
```

You can find a `ggplot` cheatsheet in the course repository (https://github.com/ErikKusch/An-Introduction-to-Biostatistics-Using-R).

How To Make A Plot In R (Basic Scatterplot)

We start off by plotting data contained within the `diamonds` data set that comes with the `ggplot2` package. We will be assessing how carats and price of individual diamonds influence each other.

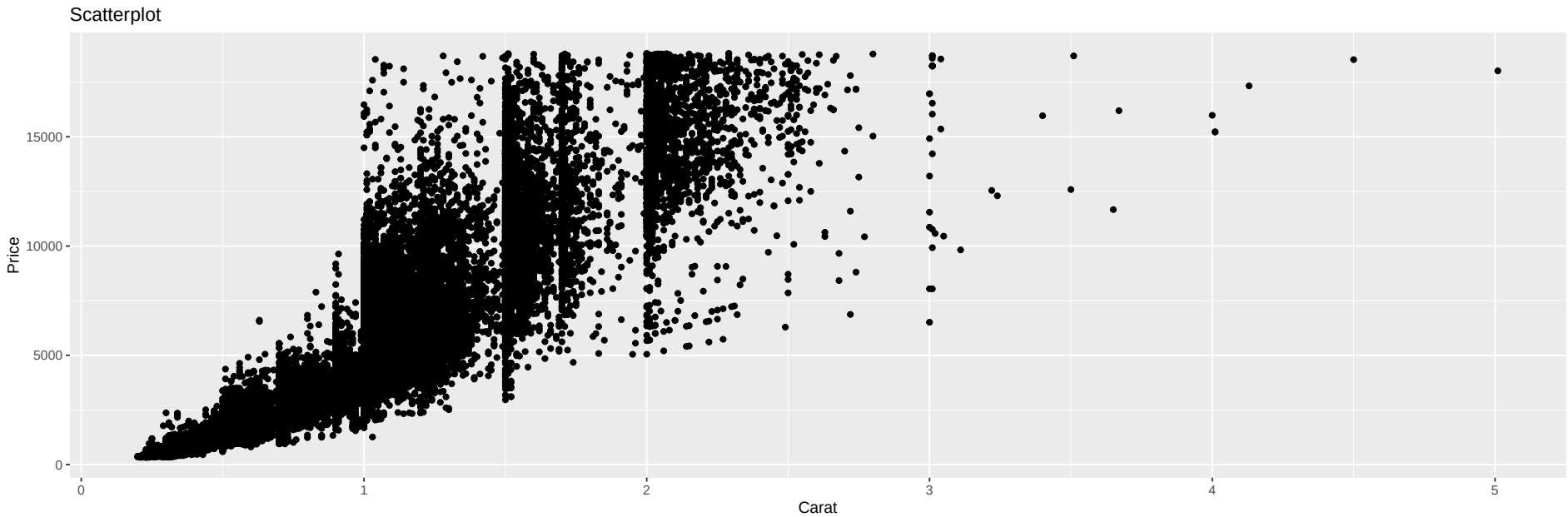
```
library(ggplot2)
p <- ggplot(diamonds, # the data set
       aes(x=carat, y=price) # aesthetics
       ) + geom_point() # geometry
p
```



How To Make A Plot In R (Labelling Axes and Title) I

A good plot always includes a title and sports some fancy axis labels:

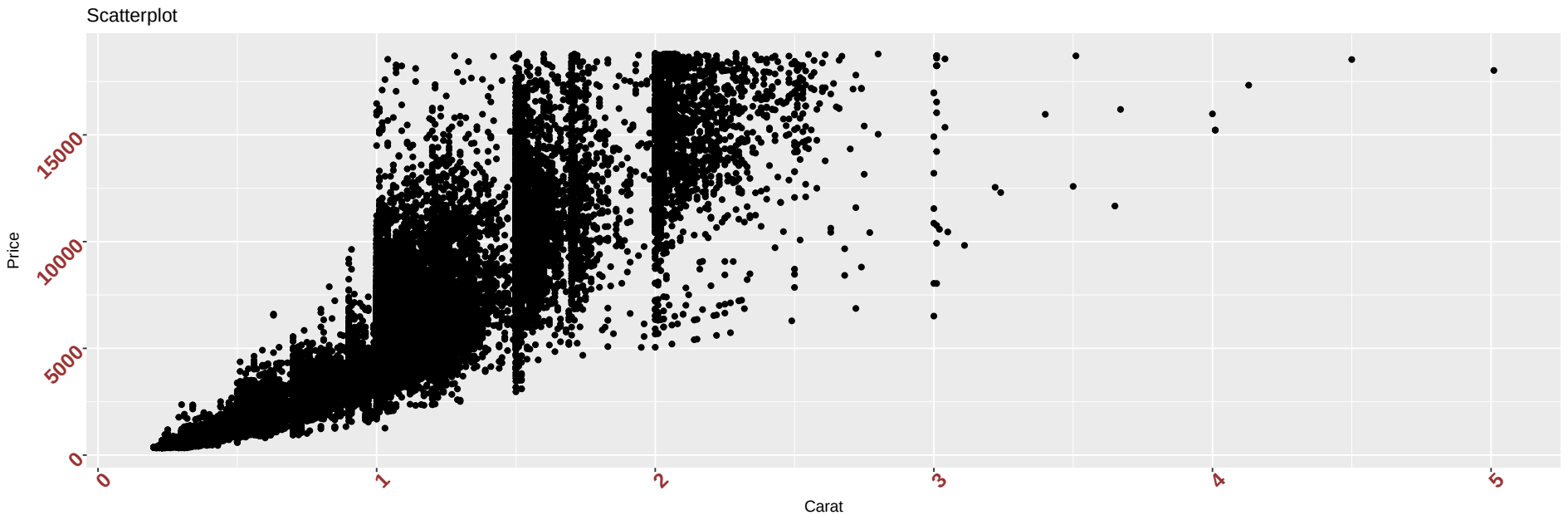
```
library(ggplot2)
p <- p + labs(title="Scatterplot", x="Carat", y="Price")
p
```



How To Make A Plot In R (Labelling Axes and Title) II

Sometimes, you may wish to customise axes even further

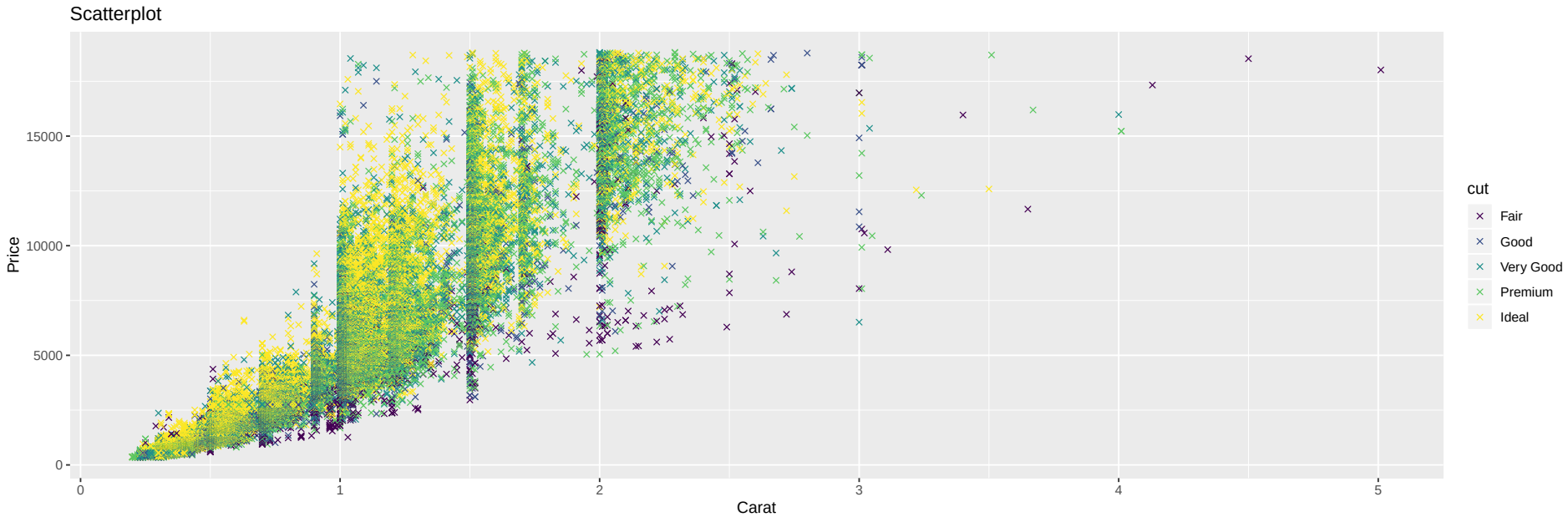
```
p <- p + theme(axis.text.x = element_text(face="bold", color="#993333",
                           size=14, angle=45),
         axis.text.y = element_text(face="bold", color="#993333",
                           size=14, angle=45))
p
```



How To Make A Plot In R (Symbols and Colours I)

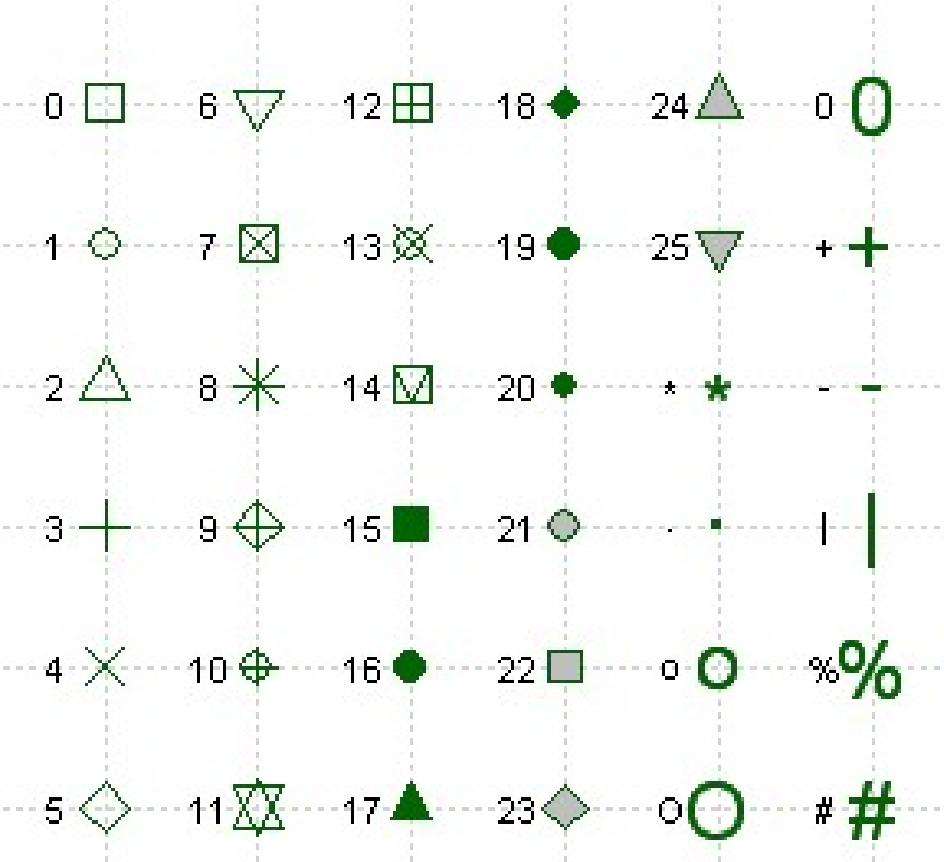
Colours are a great way of adding information to the plot. In this case, we want to visualise the quality of the cut of each diamond:

```
p <- ggplot(diamonds, aes(x=carat, y=price, color=cut)) + geom_point(shape = 4) +
  labs(title="Scatterplot", x="Carat", y="Price")
p
```



How To Make A Plot In R (Symbols and Colours II)

Symbols:



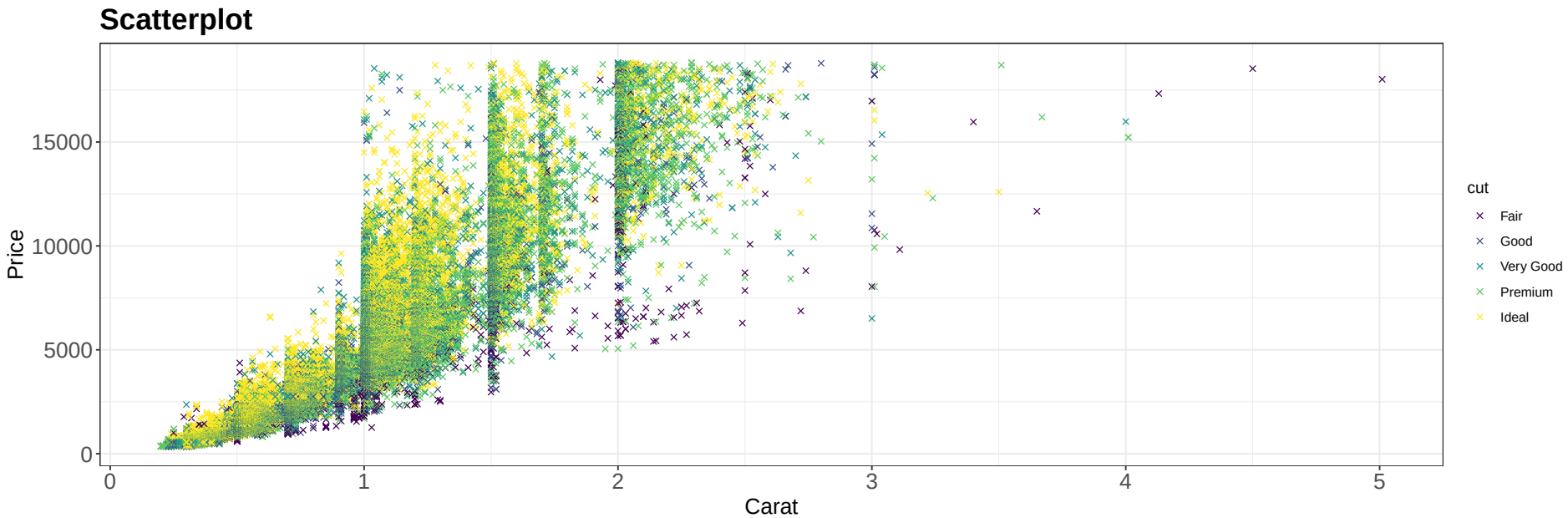
Colours:

- Hex colour codes for most precise colour specifications (https://www.color-hex.com/)
- Name specification for easiest coding (http://sape.inf.usi.ch/quick-reference/ggplot2/colour)

How To Make A Plot In R (Themes)

ggplot provides you with a set of themes for easy and quick adjustement of basic plotting components:

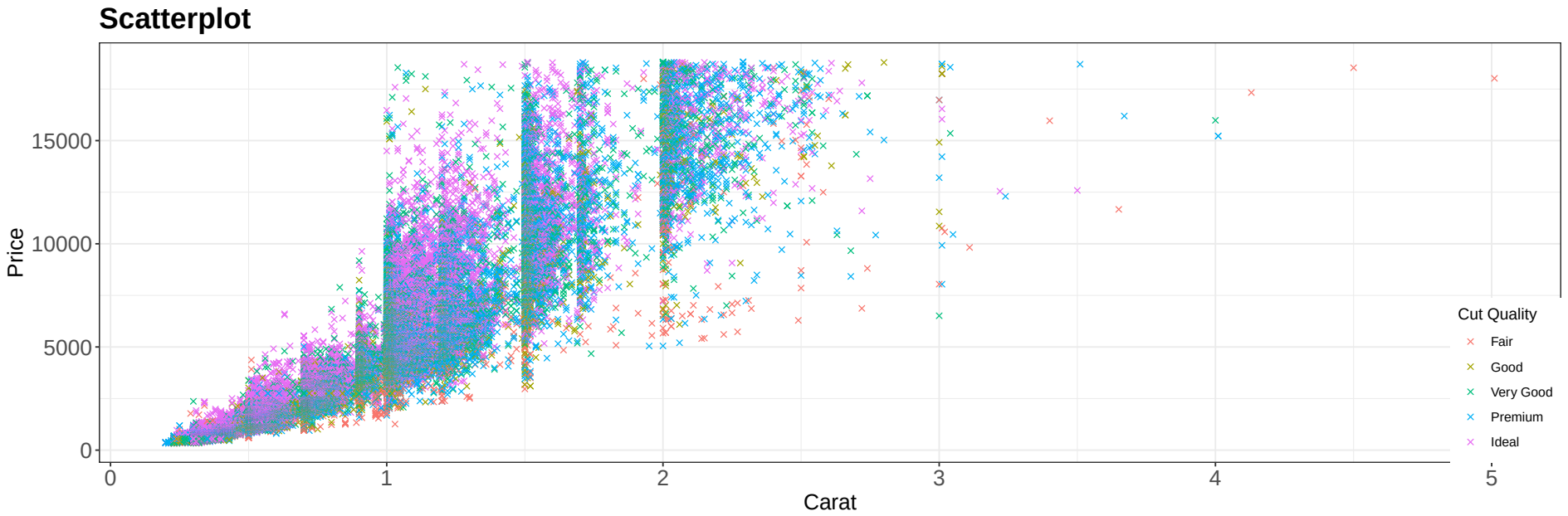
```
p <- p + theme_bw() + theme(plot.title=element_text(size=20, face="bold"),
  axis.text.x=element_text(size=15), axis.text.y=element_text(size=15),
  axis.title.x=element_text(size=15), axis.title.y=element_text(size=15))
p
```



How To Make A Plot In R (Legend)

Legends are added automatically when colours are used but may not satisfy the user:

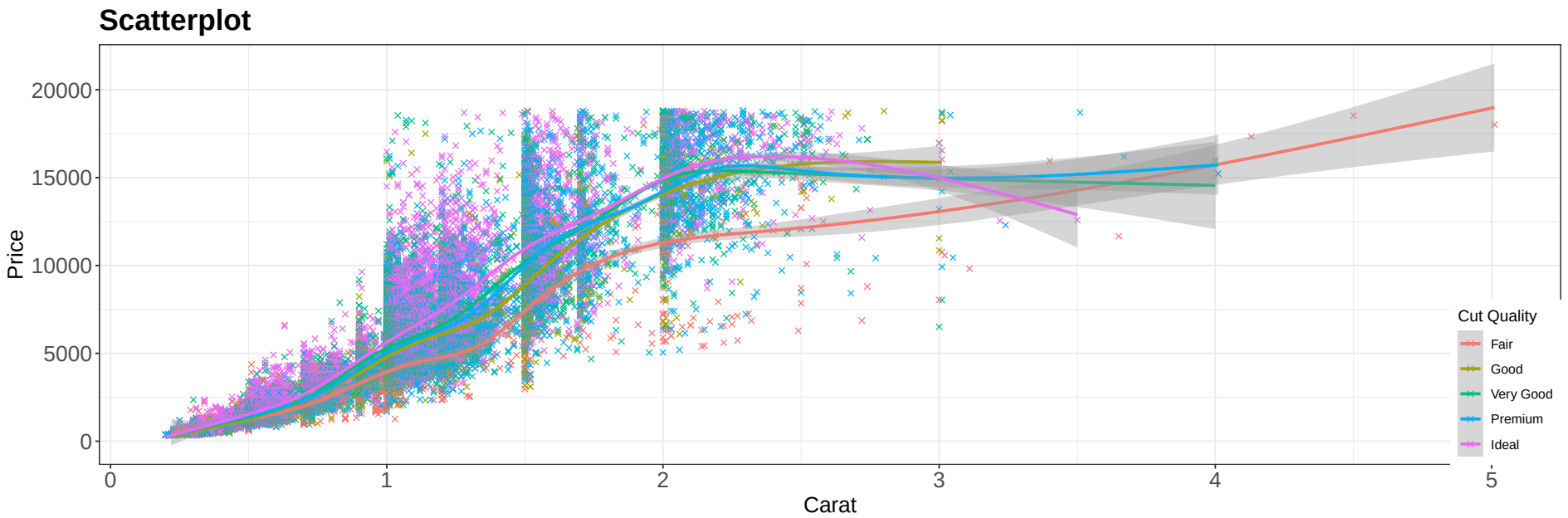
```
p <- p +
  theme(legend.justification=c(1,0), legend.position=c(1,0)) + # legend inside
  scale_color_discrete(name="Cut Quality")  # Change legend title
p
```



How To Make A Plot In R (Complex Plots)

Sometimes, you may want to show complex information that still includes the base data:

```
p <- p + geom_smooth()
p
```



How To Make A Plot In R (Saving Graphs)

Graphs can be **saved** either via the `ggsave()` function:

```
ggsave(filename = "Savedplot.jpg",
       width = 10, height = 10, units = cm)
```

or via the drop-down menu in the Files and Plots pane in RStudio.

Combining plots to appear in sets of any given number is done using the `grid.arrange()` command contained within the `gridExtra` package. For example, `grid.arrange(plot1, plot2, ncol=2)` will result in a plotting environment in which the plots (plot1 and plot 2) will be arranged side by side. These can be saved to the hard drive as follows:

```
ggsave(filename = "Savedplot.jpg",
       arrangeGrob(plot1, plot2))
```

Creating Some Data

For some of the following plotting methods, we will need the following data:

```
set.seed(42)   # making the code reproducible
data_vec <- rnorm(mean = 20, sd = 2, n = 54)
matrix(data_vec, nrow = 6)
```

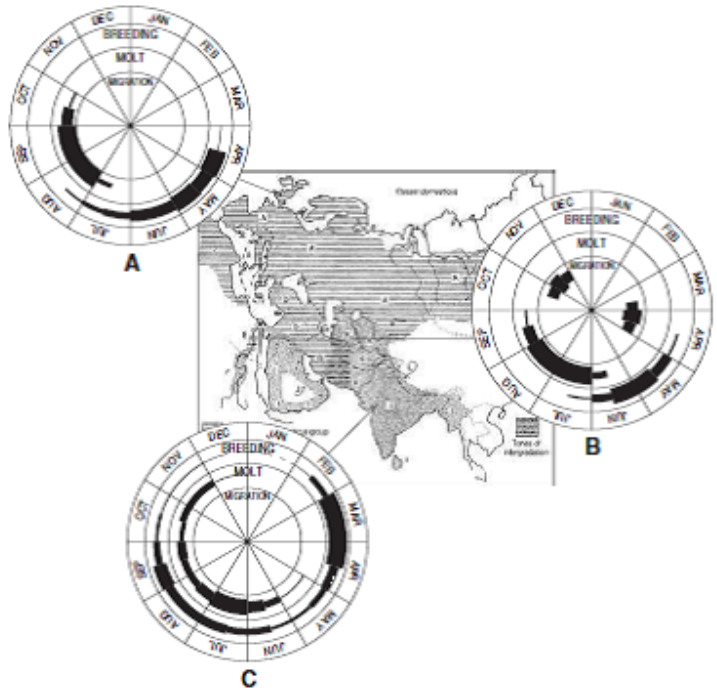
```
##          [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
## [1,] 22.74 23.02 17.22 15.12 23.79 20.91 18.43 21.52 19.14
## [2,] 18.87 19.81 19.44 22.64 19.14 21.41 18.30 18.55 21.31
## [3,] 20.73 24.04 19.73 19.39 19.49 22.07 15.17 17.26 20.64
## [4,] 21.27 19.87 21.27 16.44 16.47 18.78 20.07 20.87 18.43
## [5,] 20.81 22.61 19.43 19.66 20.92 21.01 20.41 18.38 23.15
## [6,] 19.79 24.57 14.69 22.43 18.72 16.57 19.28 22.89 21.29
```

Pie Charts In Practice

Accommodates frequency/count data.

- **For publications:**
  - Can have merit when used to show proportions.
  - Used seldom.
- **For behind-the-scenes work:**
  - Can give some initial insight on data properties.
  - Other plot types are usually preferable.

$\rightarrow$ Only really useful for showing proportions and even then line graphs may be more useful.

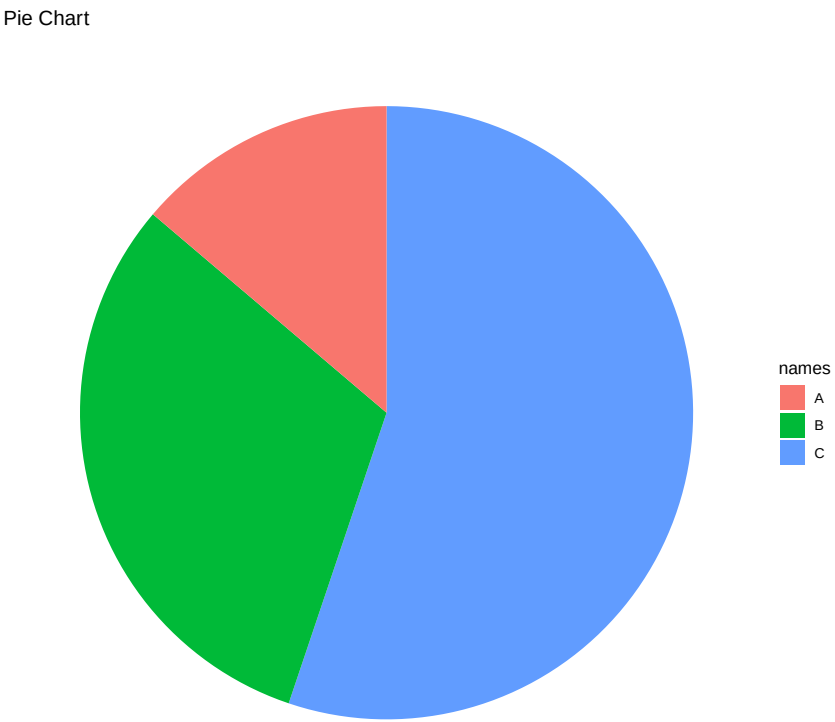


Anderson, T. (2007) Biology of the Ubiquitous House Sparrow: From Genes to Populations, Biology of the Ubiquitous House Sparrow: From Genes to Populations. doi: 10.1093/acprof:oso/9780195304114.001.0001.

Pie Charts In R

```
df <- data.frame(slices = c(4,9,16),
                 names = c("A", "B", "C"))
```

```
ggplot(df,
       aes(x="", y = slices,
           fill = names)) +
  geom_bar(width = 1,
           stat = "identity") +
  coord_polar("y", start=0) +
  theme_void() +
  labs(title = "Pie Chart")
```



Scatterplots In Practice

Accommodates all kinds of data.

- **For publications:**
  - Great way of presenting unaltered data.
  - Used extremely often.
- **For behind-the-scenes work:**
  - Perfect method for data exploration and data mining.
  - Used in almost every analysis.

→ Unavoidable data visualisation tool.

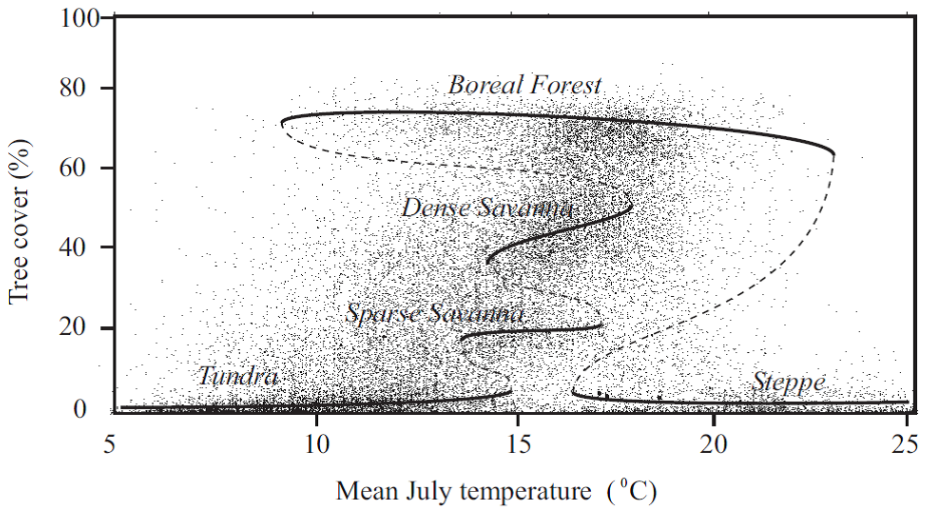


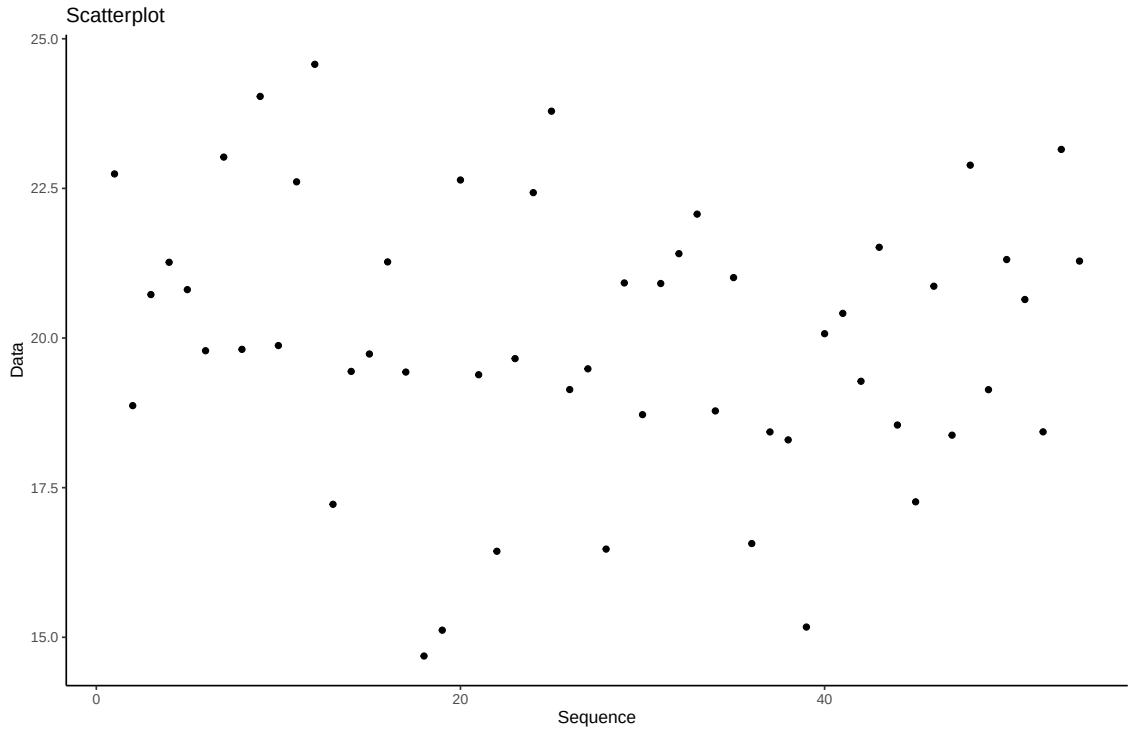
Fig. 2. Relationship between mean July temperature averaged for the period 1961–2002 and the approximate position of alternative stable states of boreal tree cover (solid curves) inferred from minima in the computed stability landscapes (*SI Appendix*, Fig. S7) computed from the data (*SI Appendix*, *SI Text* and Fig. S7). The dashed curves correspond to maxima in the computed stability landscape that separate the basins of attraction of the alternative stable states. Dots represent the tree cover and mean July temperature in the grid cells we analyzed.

Scheffer, M. et al. (2012) 'Thresholds for boreal biome transitions.', Proceedings of the National Academy of Sciences of the United States of America, 109(52), pp. 21384-9. doi: 10.1073/pnas.1219844110.

Scatterplots In R

```
df <- data.frame(
  Data = data_vec,
  Sequence = 1:length(data_vec))
```

```
ggplot(df,
       aes(x=Sequence,
           y = Data)) +
  geom_point() +
  theme_classic() +
  labs(title = "Scatterplot")
```



Line Graphs In Practice

Accommodates continuous data.

- **For publications:**
  - Often used as a logical conclusion to emerging trends in scatter plots.
  - Used pretty often. Especially when showing relationships.
- **For behind-the-scenes work:**
  - Scatter plots may suffice.
  - When causal links between variables are the goal, then these are the way to go.

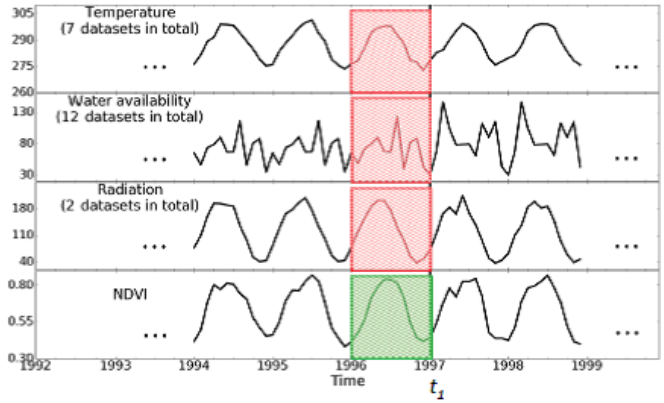


Figure 1. An illustrative example of the moving window approach considered in the analysis of vegetation drivers at a given timestamp t_1. Here, NDVI takes the role of the time series y in Eq. (3). In addition, three climate predictor time series are shown. The baseline random forest model only considers the green moving window, whereas the full random forest model includes the red moving windows as well. The pixel corresponds to a location in North America (lat: 37.5°, long: −87.5°).

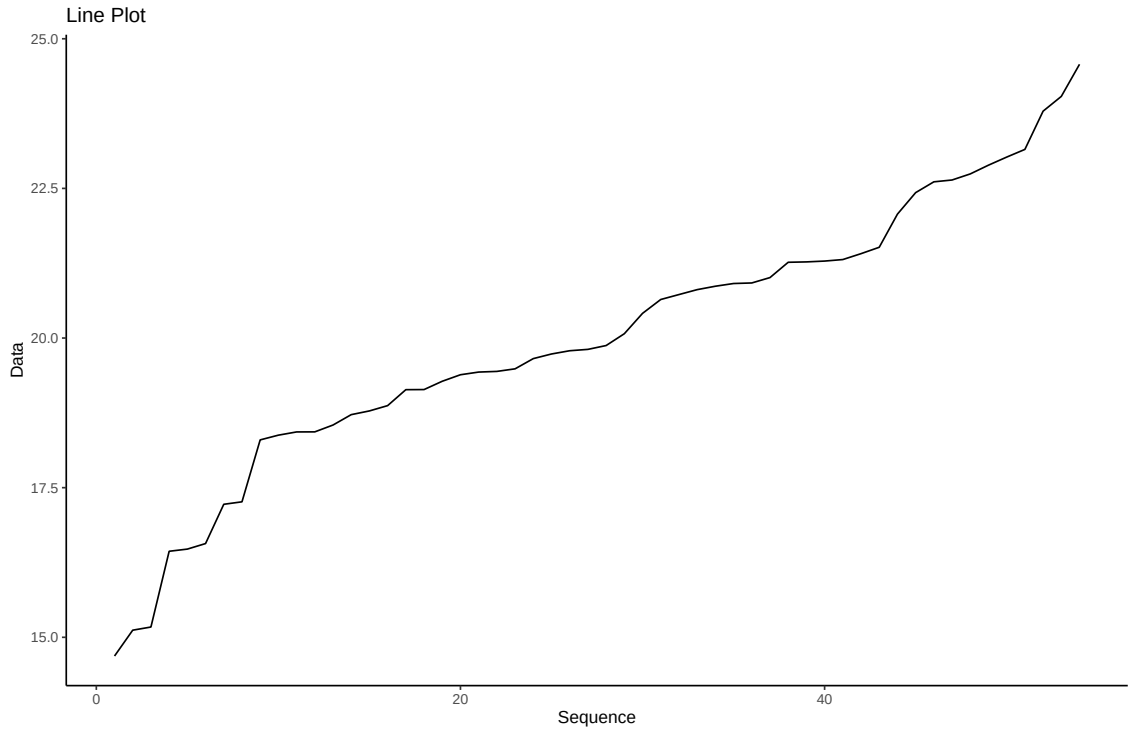
Papagiannopoulou, C. et al. (2017) 'A non-linear Granger-causality framework to investigate climate-vegetation dynamics', Geoscientific Model Development, 10(5), pp. 1945-1960. doi: 10.5194/gmd-10-1945-2017.

→ Remember only to use if continuity is actually implied

Line Graphs In R

```r
df <- data.frame(
  Data = sort(data_vec),
  Sequence = 1:length(data_vec))
```

```r
ggplot(df,
       aes(x=Sequence,
           y = Data)) +
  geom_line() +
  theme_classic() +
  labs(title = "Line Plot")
```



Bar Charts In Practice

Accommodates count data.

- **For publications:**
  - Mostly used when data can be arranged into distinct groups.
  - Used seldom.
- **For behind-the-scenes work:**
  - Can be helpful in data exploration but usually falls short of other methods.

$\rightarrow$ Useful for classifications.

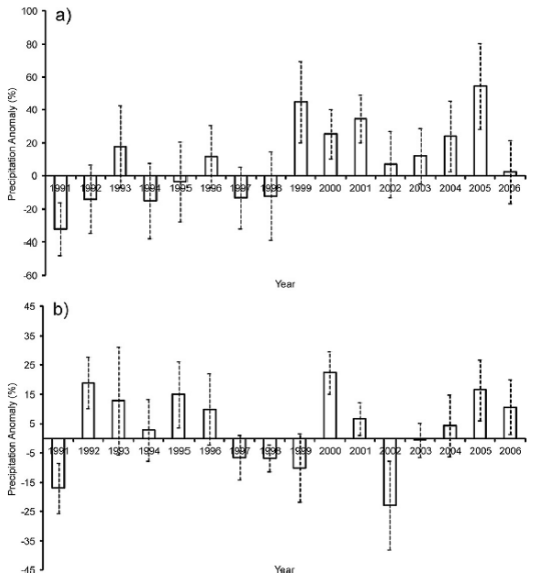


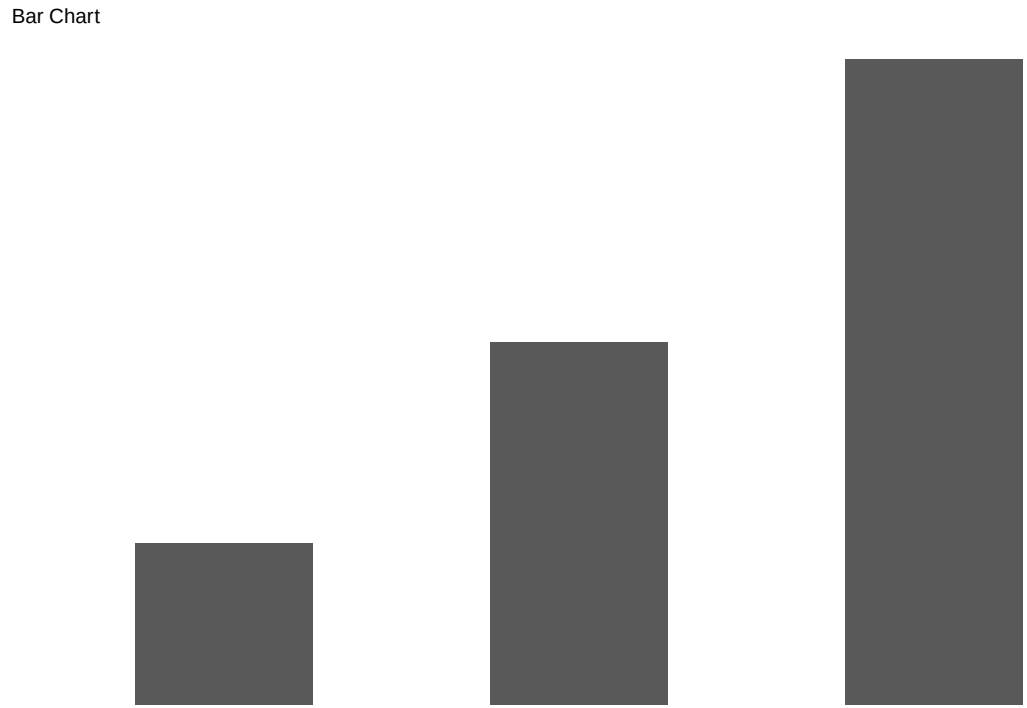
Fig. 6. Spatially aggregated precipitation anomalies derived from the CRU TS 3.0 gridded dataset for (a) the Nama Karoo and (b) the Succulent Karoo biomes. The time period 1982–1991 is taken as the reference period, against which yearly anomalies are obtained. Error bars indicate ±1 sd.

Harris, A., Carr, A. S. and Dash, J. (2014) 'Remote sensing of vegetation cover dynamics and resilience across southern Africa', International Journal of Applied Earth Observation and Geoinformation. Elsevier B.V., 28(1), pp. 131-139. doi: 10.1016/j.jag.2013.11.014.

Bar Charts In R

```
df <- data.frame(slices = c(4,9,16),
                 names = c("A", "B", "C"))
```

```
ggplot(df,
       aes(x=names,
           y = slices)) +
  geom_bar(width = .5,
           stat = "identity") +
  theme_void() +
  labs(title = "Bar Chart")
```



Histograms In Practice

Accommodates frequency count data.

- **For publications:**
  - Great way of presenting data distributions.
  - Used extensively.
- **For behind-the-scenes work:**
  - Almost unavoidable in data exploration and assumption checking.

$\rightarrow$ Used to assess and understand data distributions.

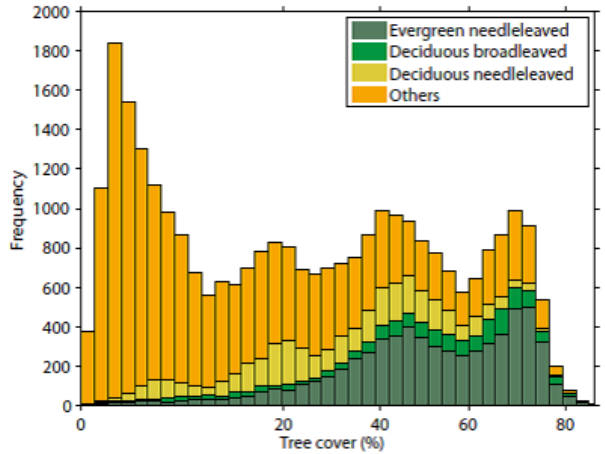
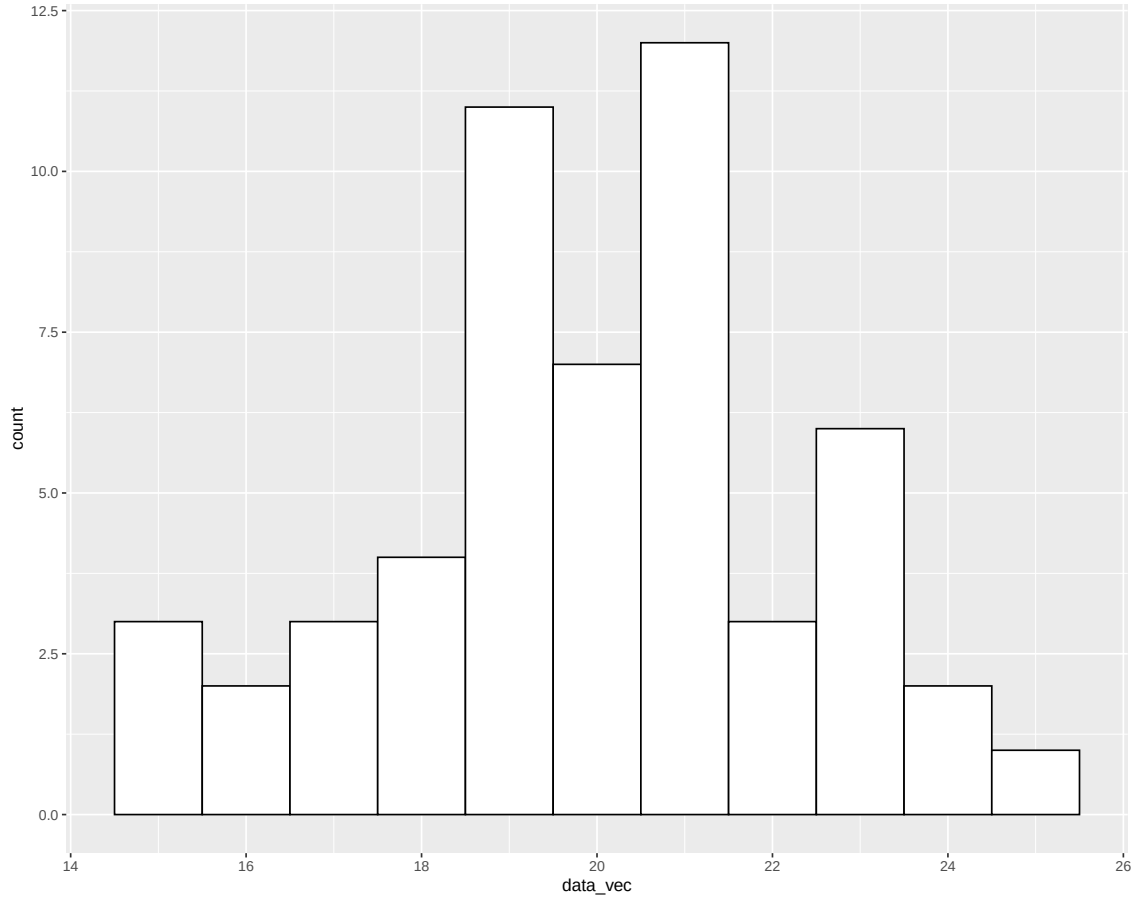


Fig. 1. Frequency distribution of tree cover in the boreal zone (45°N–70°N) in 500 × 500 m grid cells. There are four distinct modes corresponding to forest, dense savanna-like woodland, sparse savanna-like woodland, and a treeless state (tundra or steppe). Tree cover percentage values have been transformed through the arcsine-squared-root transformation.

Scheffer, M. et al. (2012) 'Thresholds for boreal biome transitions.', Proceedings of the National Academy of Sciences of the United States of America, 109(52), pp. 21384-9. doi: 10.1073/pnas.1219844110.

Histograms In R

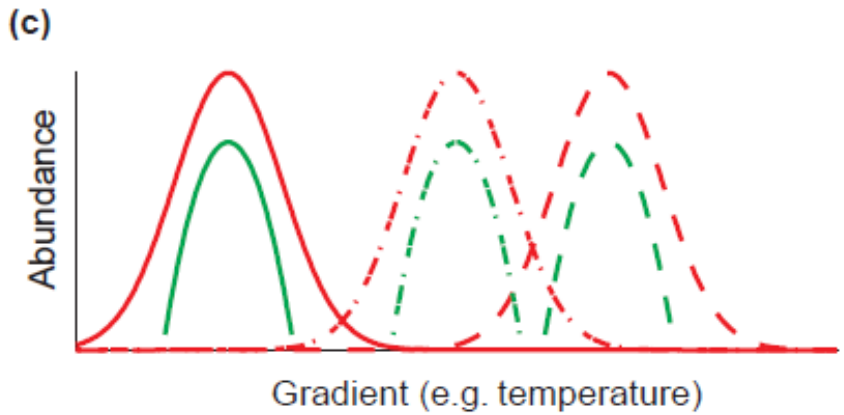
```
ggplot() + aes(data_vec)+
  geom_histogram(binwidth=1,
                 colour="black",
                 fill="white")
```



Frequency Polygon In Practice

Accommodates frequency count data.

- **For publications:**
  - May be used as the logical conclusion to histogram displays.
  - Used rather sparingly due to a possible masking effect.
- **For behind-the-scenes work:**
  - You may wish to use this to add more information to the plot besides the distribution.
  - Histograms usually suffice.

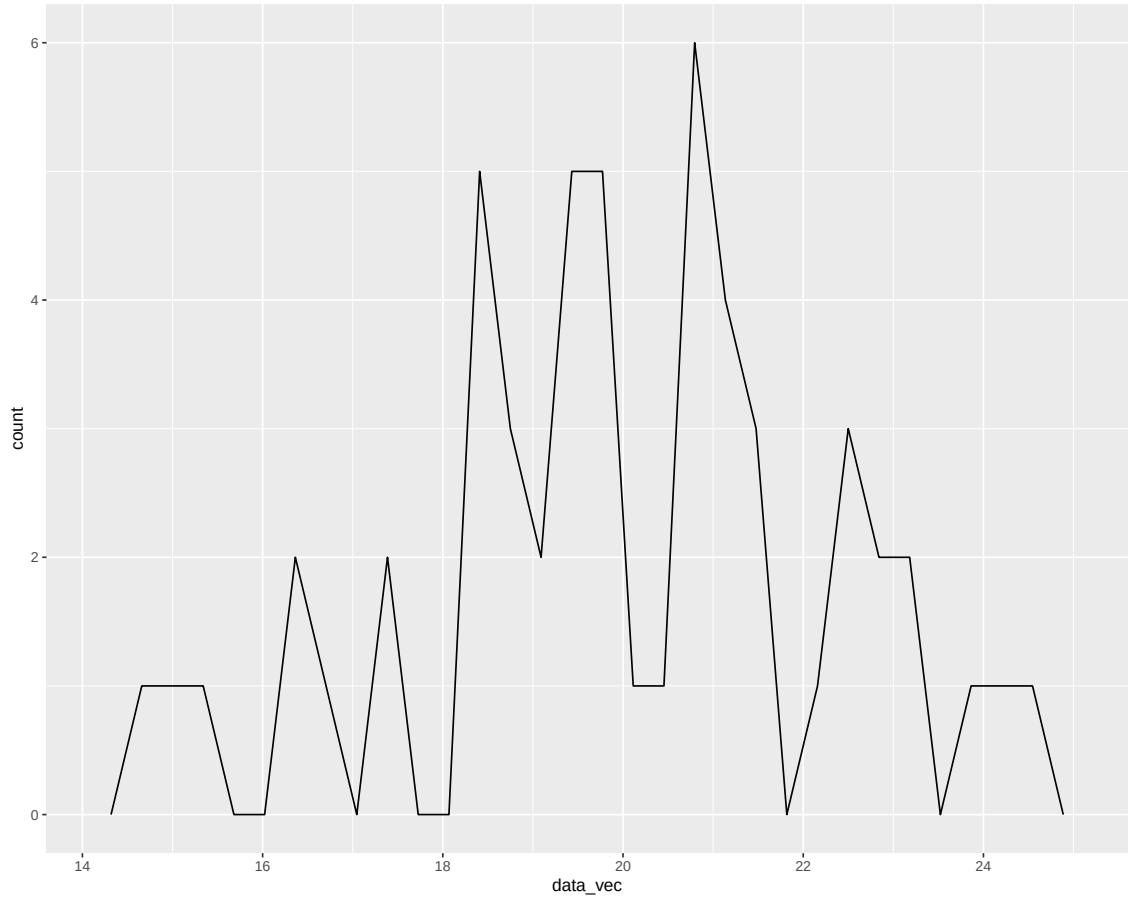


McGill, B. J. et al. (2006) 'Rebuilding community ecology from functional traits', Trends in Ecology and Evolution, 21(4), pp. 178-185. doi: 10.1016/j.tree.2006.02.002.

→ Used to assess and understand data distributions.

Frequency Polygon In R

```
ggplot() + aes(data_vec) +
  geom_freqpoly()
```



Dendrograms In Practice

Accommodates classification data.

- **For publications:**
  - Usage almost exclusively to portraying phylogenetics.
  - Applicable to all clustering approaches.
- **For behind-the-scenes work:**
  - Intuitive display of data groups.
  - Coloured scatter plots may outperform dendrograms in certain situations.

$\rightarrow$ Great to visualise hierarchical clustering approaches.

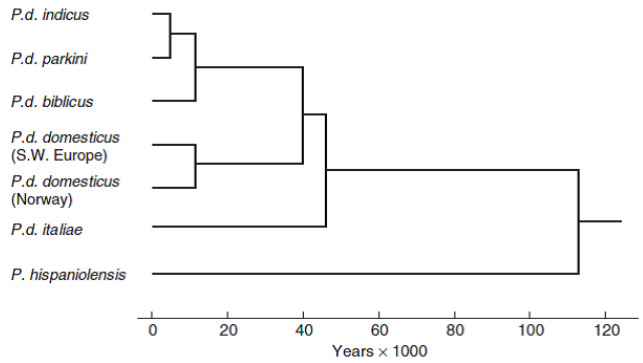


FIGURE 1.1. Dendrogram showing estimated times of divergence of several populations of the house sparrow and one population of the Spanish sparrow. Fifteen polymorphic isozyme loci (see Chapter 2) were examined in two samples each from Norway and France (*P. d. domesticus*), two samples from India (one *P. d. indicus* and one *P. d. parkini*), and one sample each from Italy (*P. d. italiae*) and Israel (*P. d. biblicus*). The one sample of the Spanish sparrow (*P. hispaniolensis*) was taken in Tunis. Note that *P. d. italiae* clusters with populations of the house sparrow rather than with the Spanish sparrow, and it has an estimated time of divergence from the latter of 113,500 ybp (Parkin 1988). Note also that *P. d. biblicus* (a member of the Palearctic group—see Fig. 1.2) clusters more closely with the two members of the Oriental group (*P. d. indicus* and *P. d. parkini*) than with the two populations of *P. d. domesticus*. From Parkin (1988: Fig. 1), courtesy of The Canadian Museum of Nature, Ottawa, Canada.

Anderson, T. (2007) Biology of the Ubiquitous House Sparrow: From Genes to Populations, Biology of the Ubiquitous House Sparrow: From Genes to Populations. doi: 10.1093/acprof:oso/9780195304114.001.0001.

Dendrograms In R

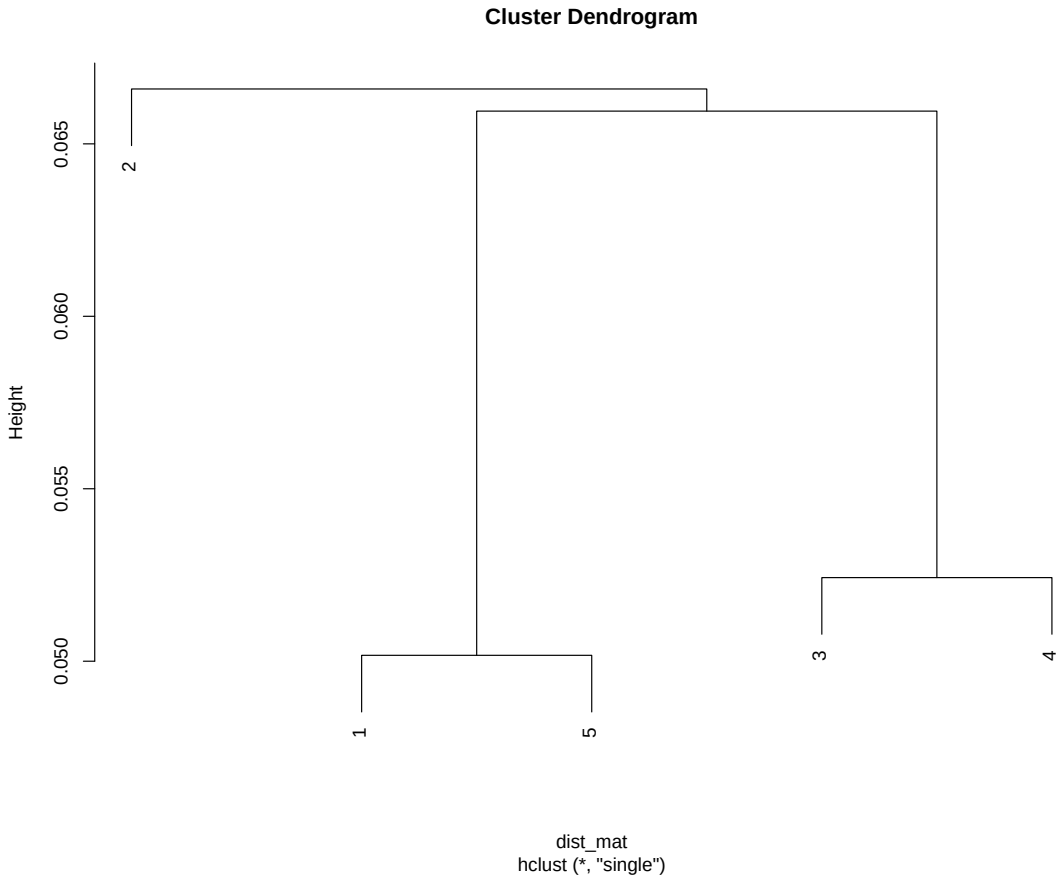
`ggplot` can't handle certain objects (such as these hierarchical clusters):

```
library(vegan)

dist_mat <- vegdist(
  matrix(data_vec[1:25],
         nrow=5))

clust <- hclust(
  d = dist_mat,
  method="single")

plot(clust)
```



Boxplots In Practice

Accommodates numerical data.

- **For publications:**
  - Immensely useful data visualisation tool to represent parameters of groups of data.
  - Used very frequently.
- **For behind-the-scenes work:**
  - Always nice for data exploration.
  - Hard to avoid (not that you'd want to).

$\rightarrow$ Used to present basic parameters of descriptive statistics.

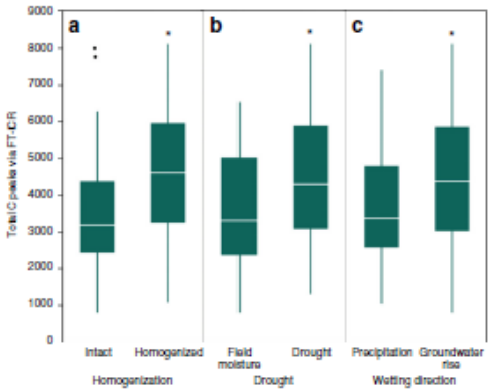


Fig. 3 Total C peaks identified in soil pore waters by soil homogenization, antecedent drought and wetting direction. The total number of Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometry peaks of organic C identified across all pore water fractions that significantly differed by **a** soil homogenization ($P = 0.001$, $n = 43$ for intact, 42 for homogenized), **b** antecedent drought ($P = 0.011$, $n = 46$ for field moisture, 39 for antecedent drought), and **c** wetting direction ($P = 0.034$, $n = 44$ for simulated precipitation, 41 for simulated groundwater rise). The outlier box plot whiskers represent the first and third quartile minus or plus, respectively, 1.5 times the interquartile range. Soil pore water was collected immediately following rewetting and post-rewetting incubation

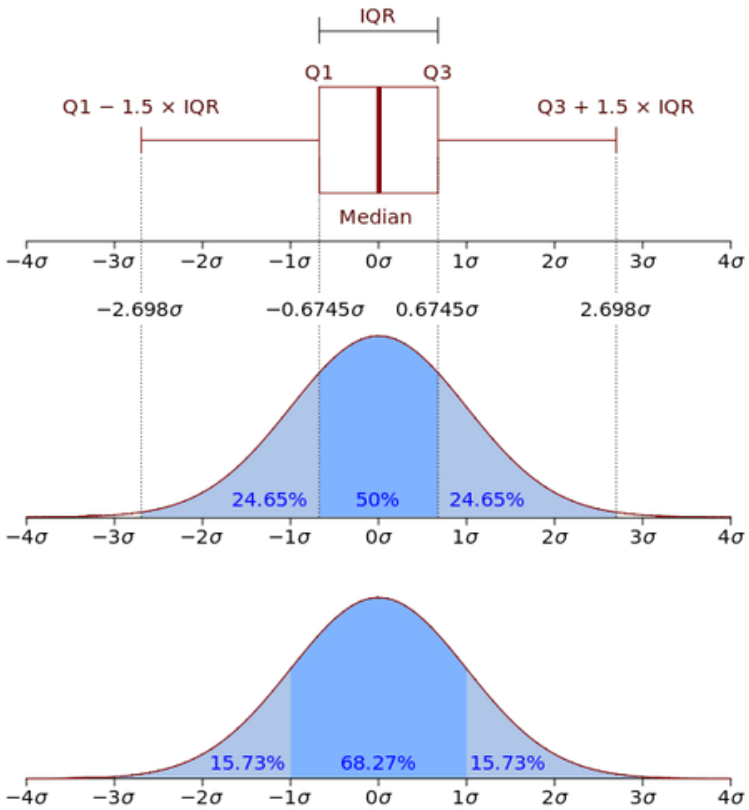
Smith, A. P. et al. (2017) 'Shifts in pore connectivity from precipitation versus groundwater rewetting increases soil carbon loss after drought', Nature Communications. Springer US, 8(1), p. 1335. doi: 10.1038/s41467-017-01320-x.

Boxplots In Theory

Box plots are less intuitive than other plotting displays:

Contained information:

- Lower and upper 99.3% intervals of the data (expressed as whiskers).
- The cut-point for Quartile 1 and 3 (these are the outer edges of the box, so 50% of the data fall inside the box).
- The Median, usually represented by a bold line inside the box, because its behaviour is robust (more so than that of the mean).



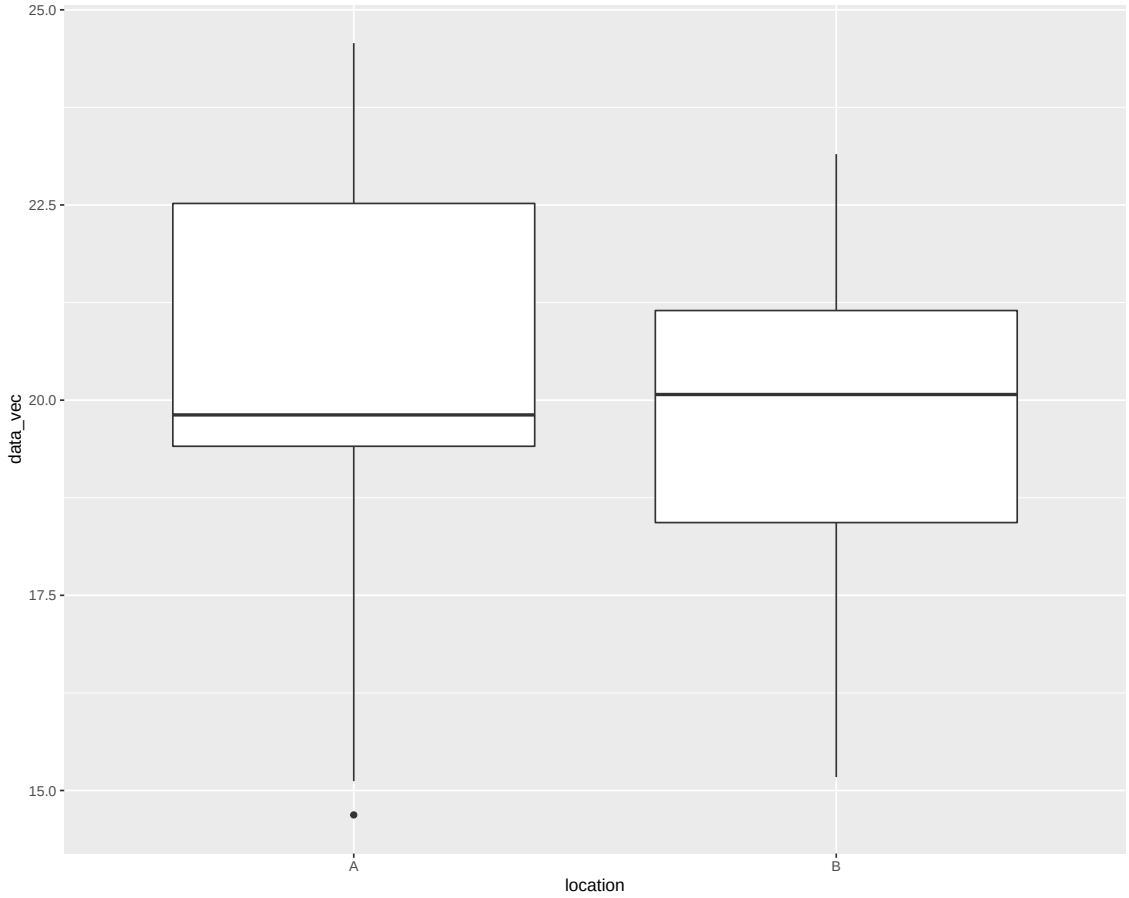
Boxplots In R

You can also use `geom_violin()` for some fancy violin plots which result in a roughly equal depiction of the data.

```
location <- as.factor(
  c(rep("A",27),
    rep("B",27)))

data_df <- data.frame(
  data_vec,location)

ggplot(data_df,
       aes(x = location,
           y = data_vec)) +
  geom_boxplot()
```

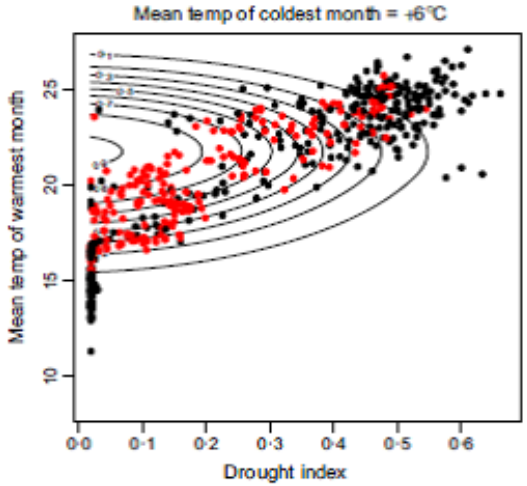


Contour Plots In Practice

Accommodates all kinds of data.

- **For publications:**
  - More complicated to understand.
  - Used sparingly.
- **For behind-the-scenes work:**
  - You might as well include it in your final manuscript if you bother coming up with one.

  $\rightarrow$ Used to understanding the relationship of variables in a classification setting.



Brewer, M. J. et al. (2016) 'Plateau: A new method for ecologically plausible climate envelopes for species distribution modelling', Methods in Ecology and Evolution, pp. 1489-1502. doi: 10.1111/2041-210X.12609.

3-D Plots In Practice

Accommodates all kinds of data.

- **For publications:**
  - $>$ 2 dimensions translate badly to paper.
  - Used extremely rarely.
- **For behind-the-scenes work:**
  - Good for data exploration.
  - Especially useful when inspecting PCA (Principal Component Analysis) results.

$\rightarrow$ Used to understanding the relationship of variables in a classification setting.

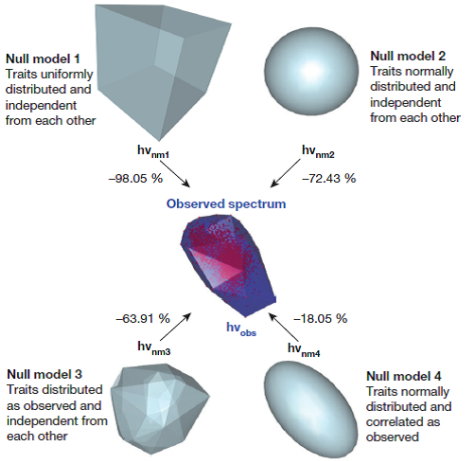


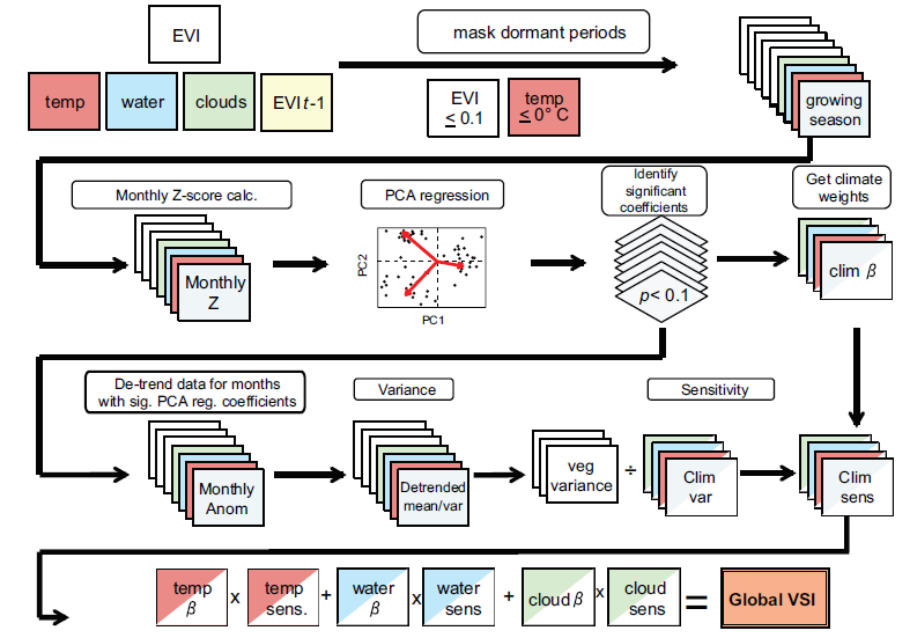
Figure 1 | The volume in trait space occupied by vascular plant species is strongly constrained compared to theoretical null models. The five diagrams are pictorial representations based on three out of the six trait dimensions forming the hypervolumes under scrutiny. The hypervolumes are constructed on the basis of $\log_{10}$- and z-transformed observed values of H, SSD, LA, LMA, N_{mass} and SM (observed hypervolume = hv_{obs}), or on the bases of four different null models of multivariate variation of those traits (hv_{nm1} to hv_{nm4}) (see Methods). Numbers adjacent to arrows indicate percentage reductions in size of hv_{obs} compared to the null-model hypervolumes (all significant at $P < 0.001$).

Díaz, S. et al. (2015) 'The global spectrum of plant form and function', Nature. Nature Publishing Group, 529(7585), pp. 167-171. doi: 10.1038/nature16489.

The Hat Goes Deeper!

There are way more plot types that you may want to use at some point.

Flow charts to illustrate your workflow, for example:



Extended Data Figure 1 | Study Design. Flow chart of the algorithm used to estimate the vegetation sensitivity index.

Seddon, A. W. R. et al. (2016) 'Sensitivity of global terrestrial ecosystems to climate variability.', Nature, 531(7593), pp. 229-232. Available at: http://dx.doi.org/10.1038/nature16986.

The data that comes with R

R is supplied with in-built data sets and more data sets will be added to your local library when you install additional packages. These data sets are immensely useful in creating **minimal working exmaples (MWEs)** which show how something works (or doesn't) with the least amount of code possible.

You can **retrieve all available data** sets in your library using the command `data()` and **load** any of the given **data sets** by adding the name of the data set as the argument to the `data()` function.

Creating plots with R

Your **ToDo-List** for this exercise:

- Load the R-internal `iris` data set (it is included in the `datasets` package)
- Inspect the data set
- Produce a **boxplot** of `Petal.Length` by `Species`
- Produce a **scatterplot** of `Petal.Length` and `Petal.Width`
- Produce a **scatterplot** of `Petal.Length` and `Petal.Width` grouped by `Species`
- Produce a **plot of your choice** to show the relationship of `Sepal.Length` and `Sepal.Width`
- Produce a **plot of your choice** to show the relationship of `Sepal.Length` and `Sepal.Width` when grouped by `Species`

→ Play around with other combinations of variables and plotting types