ETC5521: Exploratory Data Analysis

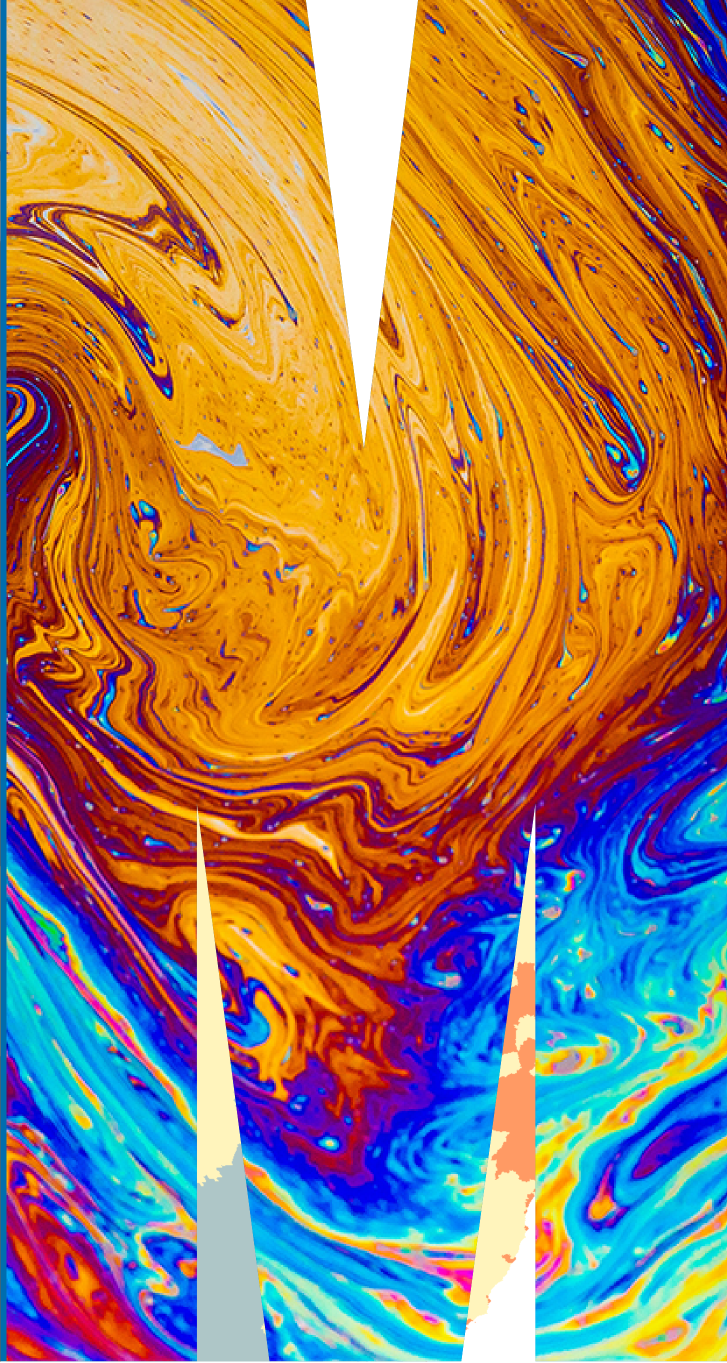
Exploring data having a space and time context

Lecturer: *Di Cook*

Department of Econometrics and Business Statistics

✉ ETC5521.Clayton-x@monash.edu

Week 10 - Session 1



> *You show me continents, I see the islands, You count the centuries, I blink my eyes*

Björk

Outline

- Breaking up data by time, and by space
- Maps of space over time
- Exploring time over space glyph maps
- Bending the choropleth map
- A flash back to the 1970s: Tukey's median polish

Case study ① Temperature change in Americas

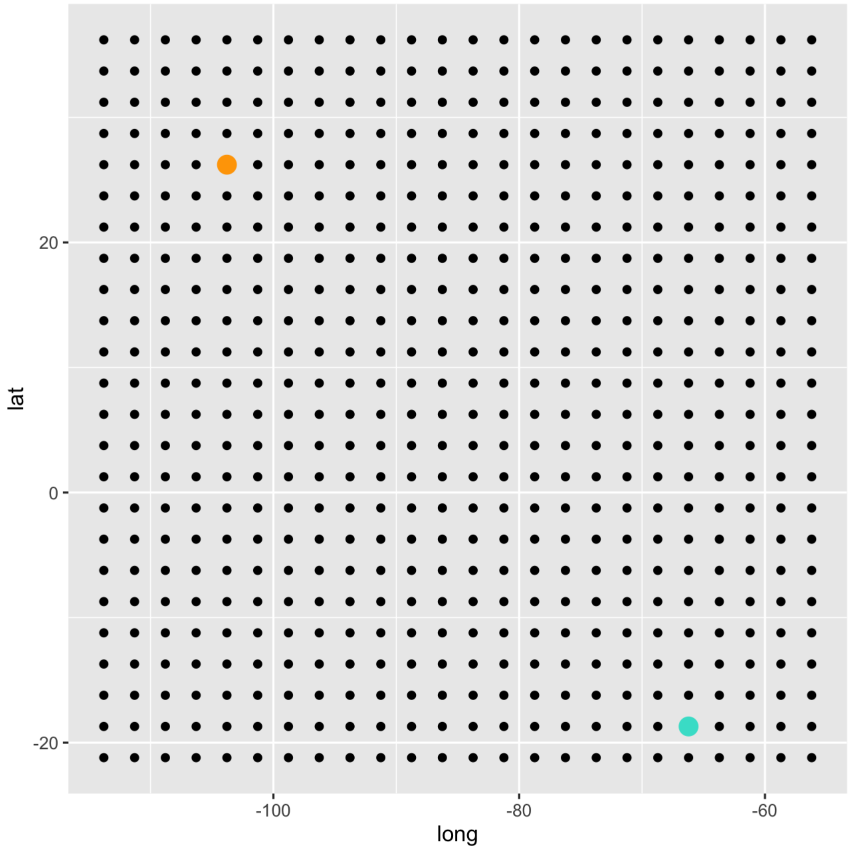
`data` R

6 years of monthly measurements of a 24x24 spatial grid from Central America collated by Paul Murrell, U. Auckland.

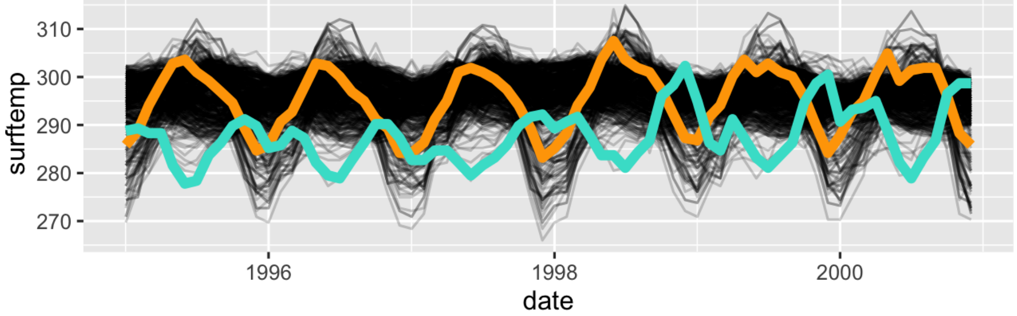
time	y	x	lat	long	date	cloudhigh	cloudlow	cloudmid	ozone	pressure	su
1	1	1	-21.2	-113.80000	1995-01-01	0.5	31.0	2.0	260	1000	29
1	1	2	-21.2	-111.29565	1995-01-01	1.5	31.5	2.5	260	1000	29
1	1	3	-21.2	-108.79130	1995-01-01	1.5	32.5	3.5	260	1000	29
1	1	4	-21.2	-106.28696	1995-01-01	1.0	39.0	4.0	258	1000	29
1	1	5	-21.2	-103.78261	1995-01-01	0.5	48.0	4.5	258	1000	29

Case study 1 Temperature change in Americas

plot R



plot R

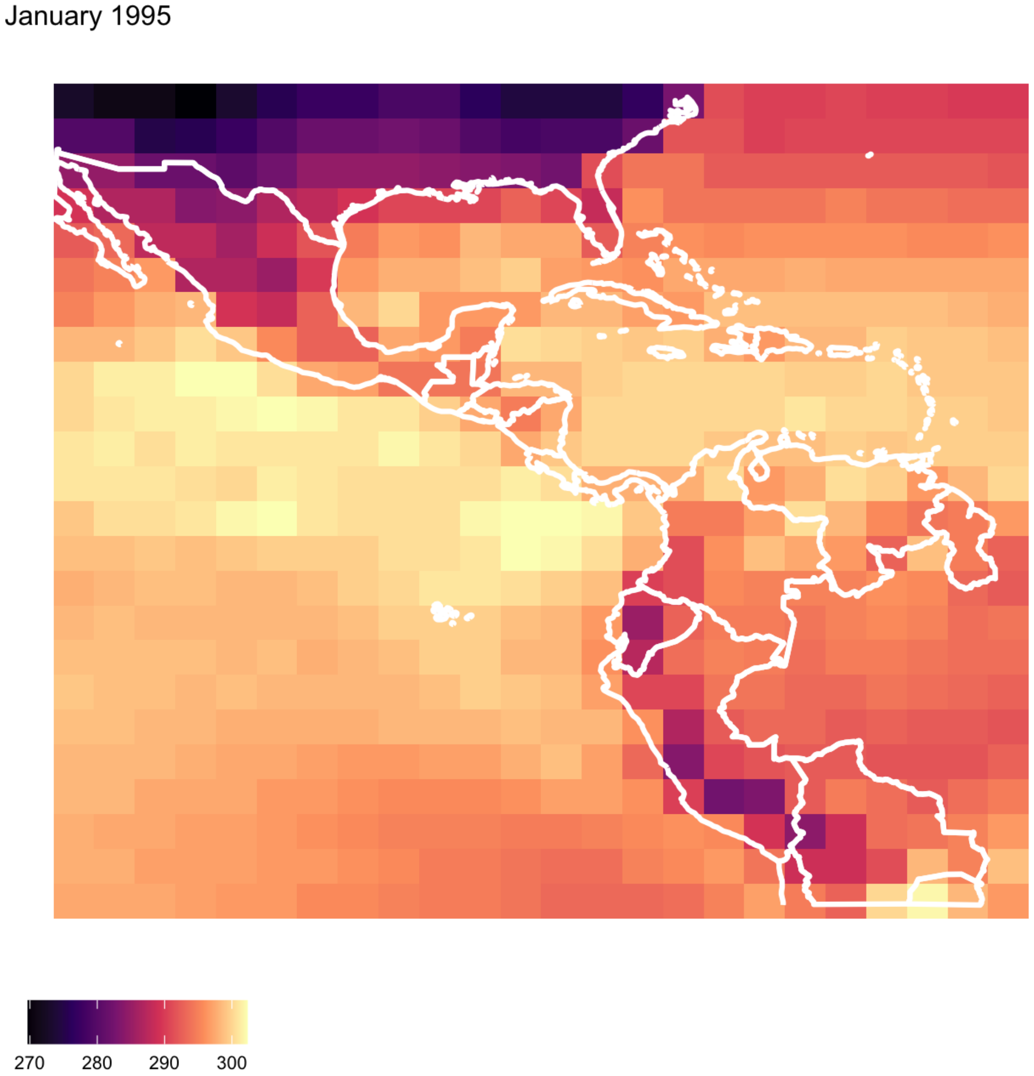


Pre-processing of time and space

Think of time and space as a categorical variables. You may need to create the categories of time. Spatial variable might need to be discretised, or gridded.

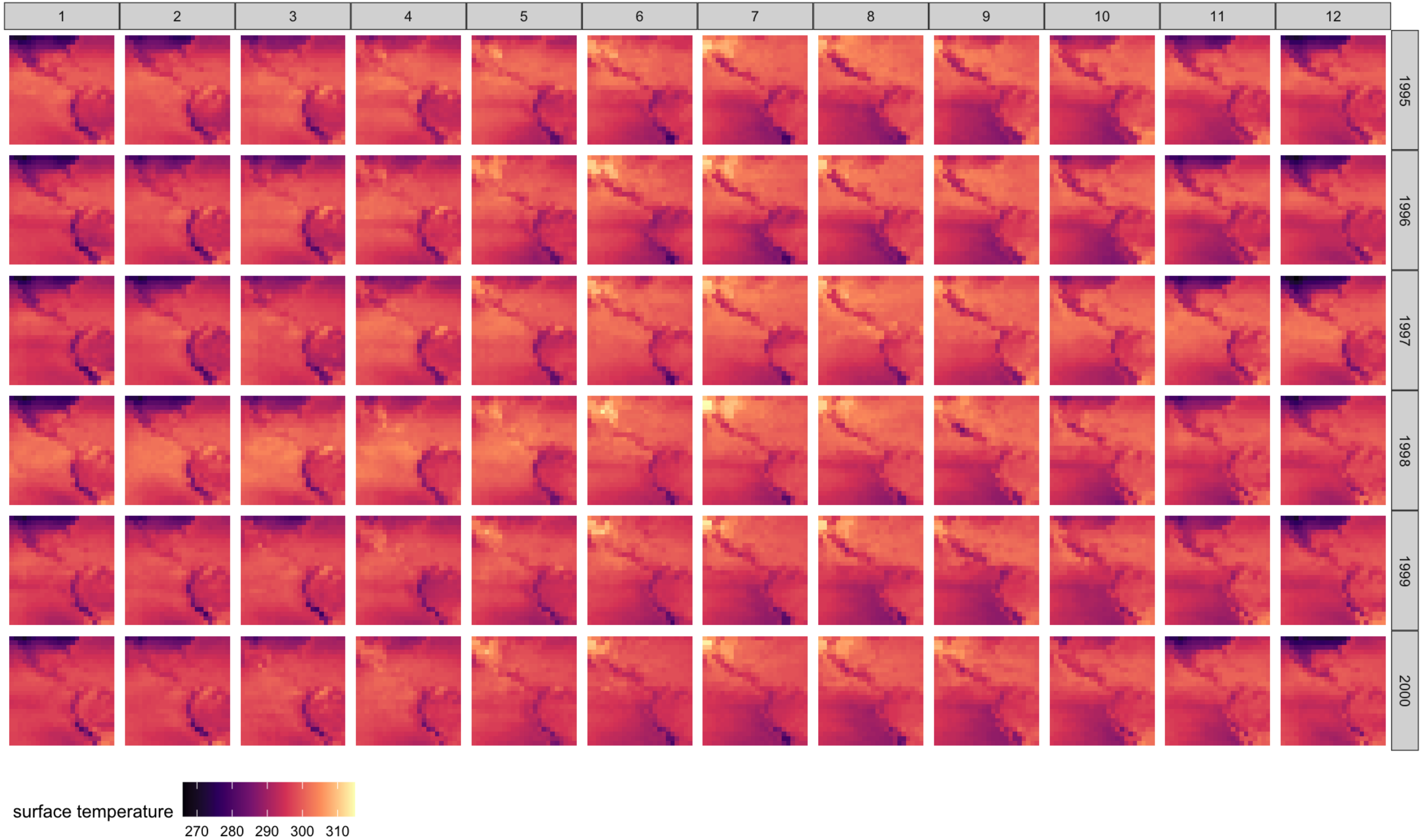
Slice in time and create a spatial map

`plot` `learn` `R`



Expand using facets in time

`plot` `learn` `R`



Expand time across space

`plot` R

This is called a glyph map. Small time series are plotted at each spatial location.

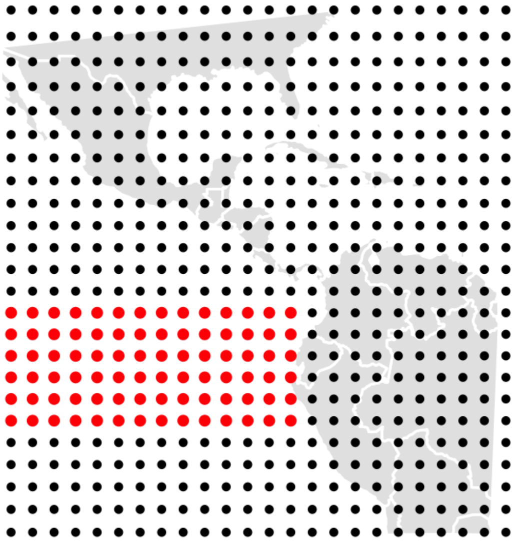
`plot` R

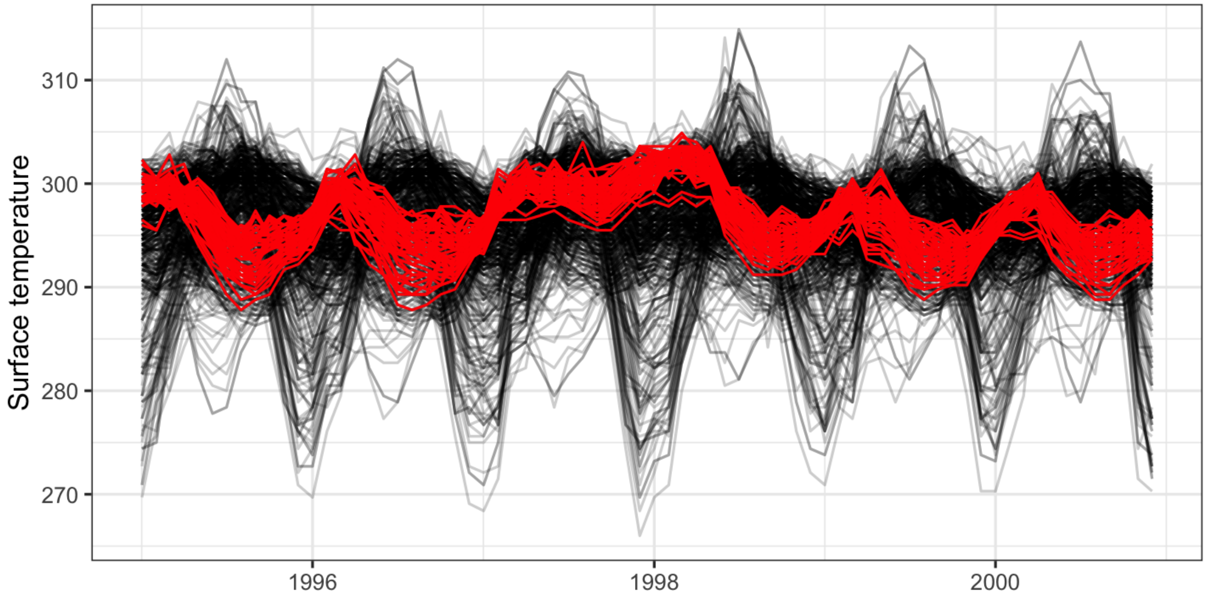
This is for exploring temporal trends over space. Here the time series are represented in polar coordinates.

Case study ① Temperature change in Americas

Detecting El Nino

Slice space, and show the time series, and the pattern is very clear: The seasonal water temperature decrease doesn't happen in 1997, and water in this area stays unseasonably warm.





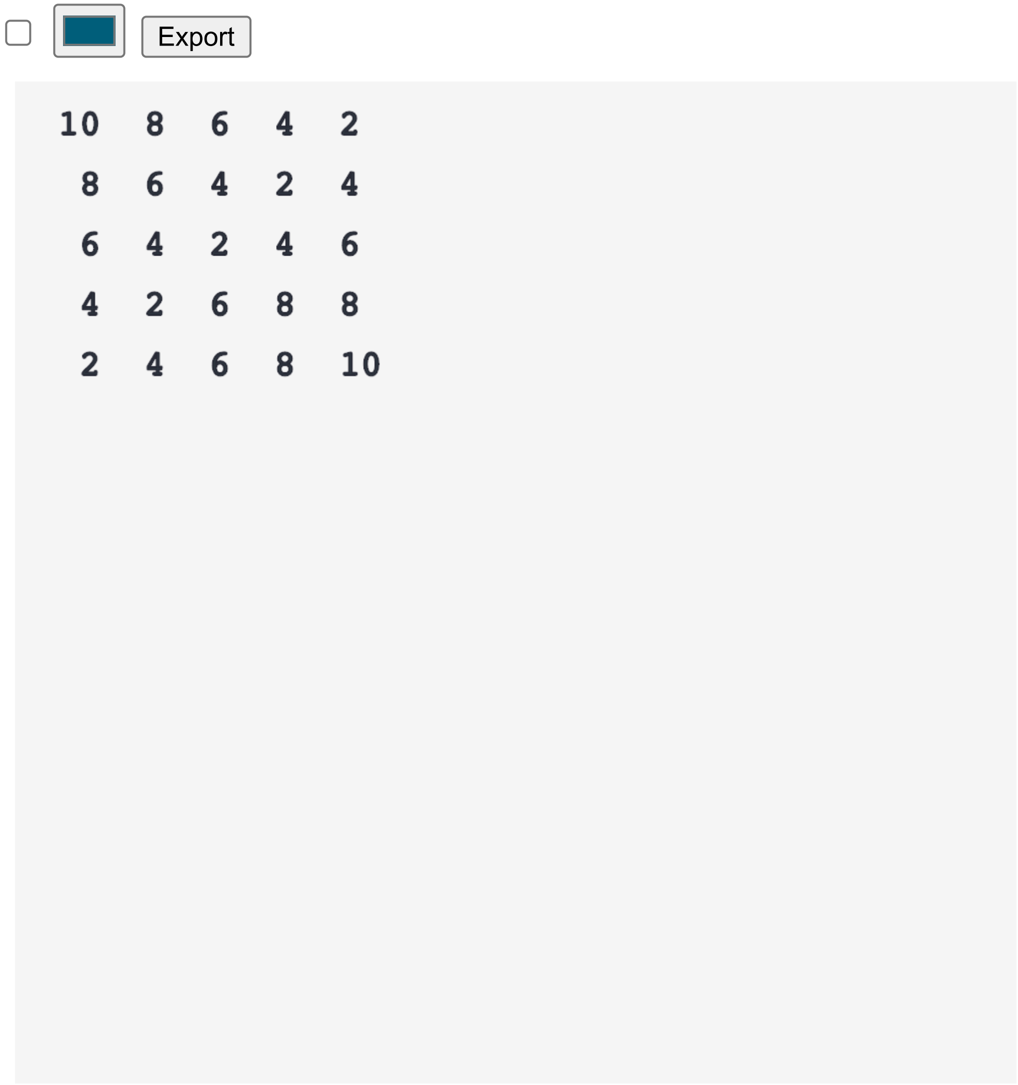
🔧 Your turn using tsibbletalk

```r
library(tsibble)
library(tsibbletalk)
library(lubridate)
nasa_shared <- nasa %>%
  mutate(date = ymd(date)) %>%
  select(long, lat, date, surftemp, id) %>%
  as_tsibble(index=date, key=id) %>%
  as_shared_tsibble()
p1 <- nasa_shared %>%
  ggplot(aes(x = long, y = lat)) +
  geom_point(aes(group = id))
p2 <- nasa_shared %>%
  ggplot(aes(x = date, y = surftemp)) +
  geom_line(aes(group = id), alpha = 0.5)
library(plotly)
subplot(
    ggplotly(p1, tooltip = "Region", width = 100),
    ggplotly(p2, tooltip = "Region", width = 900),
    nrows = 1, widths=c(0.4, 0.6)) %>%
  highlight(dynamic = TRUE)
```

A flash back to the 1970s: Tukey's median polish

This is a useful data scratching technique, particularly for spatial data, to remove complicated trends.

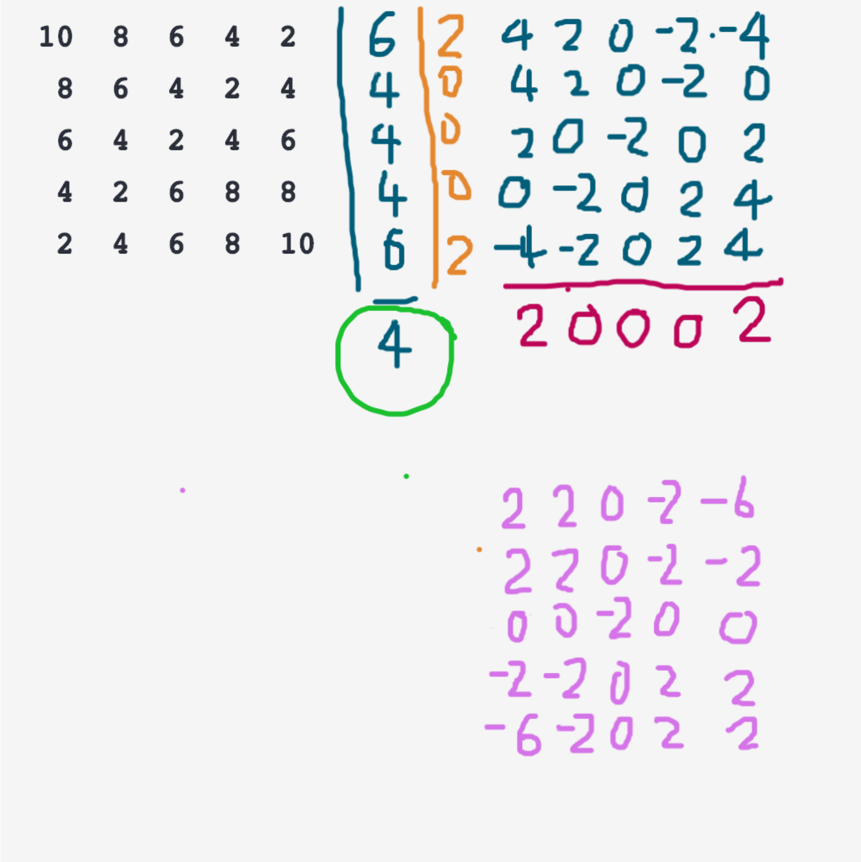
Median polish technique



```
10   8   6   4   2
 8   6   4   2   4
 6   4   2   4   6
 4   2   6   8   8
 2   4   6   8  10
```

1. Compute row medians, and the median of the row medians, called **row overall effect**.
2. Subtract each element in a row by its row median.
3. Subtract the row overall effect from each row median.
4. Do the same columns. Add the column overall effect to row overall effect.
5. Repeat 1-4 until negligible change occur with row or column medians.

Median polish technique



10	8	6	4	2	6	2	4	2	0	-2	-4
8	6	4	2	4	4	0	4	2	0	-2	0
6	4	2	4	6	4	0	2	0	-2	0	2
4	2	6	8	8	4	0	0	-2	0	2	4
2	4	6	8	10	6	2	-4	-2	0	2	4
					4		2	0	0	0	2

2	2	0	-2	-6
2	2	0	-2	-2
0	0	-2	0	0
-2	-2	0	2	2
-6	-2	0	2	2

```
## 1: 42


##
## Median Polish Results (Dataset: "x")
##
## Overall: 4
##
## Row Effects:
## [1] 2 0 0 0 2
##
## Column Effects:
## [1] 2 0 0 0 2
##
## Residuals:
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    2    2    0   -2   -6
## [2,]    2    2    0   -2   -2
## [3,]    0    0   -2    0    0
```

Median polish technique

```
## 1: 42
## Final: 42

##
## Median Polish Results (Dataset: "x")
##
## Overall: 4
##
## Row Effects:
## [1] 2 0 0 0 2
##
## Column Effects:
## [1] 2 0 0 0 2
##
## Residuals:
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    2    2    0   -2   -6
## [2,]    2    2    0   -2   -2
```

Median polish is effectively fitting a model of this form:

overall effect + row effect + column effect

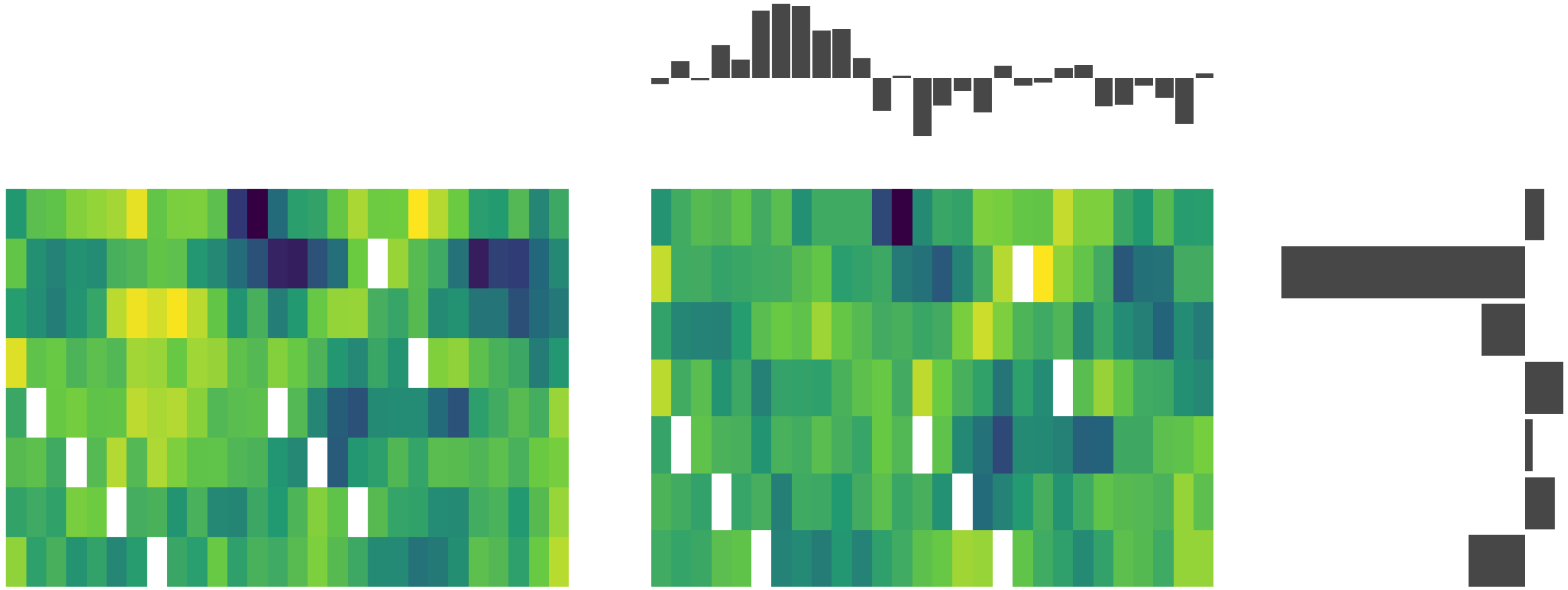
which can be written as:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

Nice explanation by Manny Gimond (2020)

Case study 2 Soils

`plot   R`



This is the baker field data that we have seen before. The heatmap shows corn yield in a farm field in Iowa. High values are yellow and low values are dark blue.

The right-side heatmap shows the residuals from median polish, and the row and column marginal effects. After a median polish, the values should look randomly distributed.

Choropleth maps and cartograms

Choropleth maps

A choropleth map is used to show a measured variable associated with a political or geographic region. Polygons for the region are filled with colour.

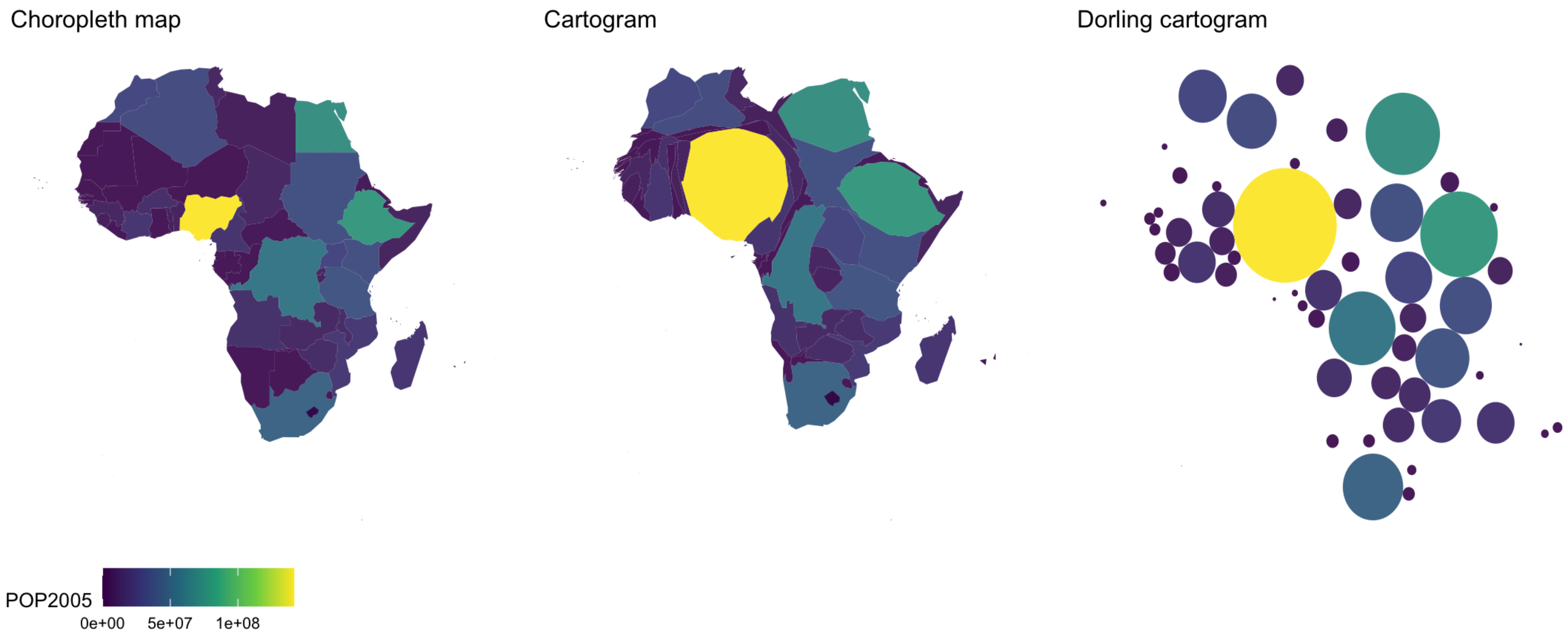
Choropleth maps

The problem with choropleth maps is that geographically large areas dominate the view and obscure the statistics of small regions.

Cartograms

A cartogram transforms the geographic shape to match the value of a statistic. Its a useful exploratory technique for examining the spatial distribution of a measured variable.

`plot   R`



That's it, for this lecture!

Lecturer: Di Cook

Department of Econometrics and Business Statistics

✉ ETC5521.Clayton-x@monash.edu