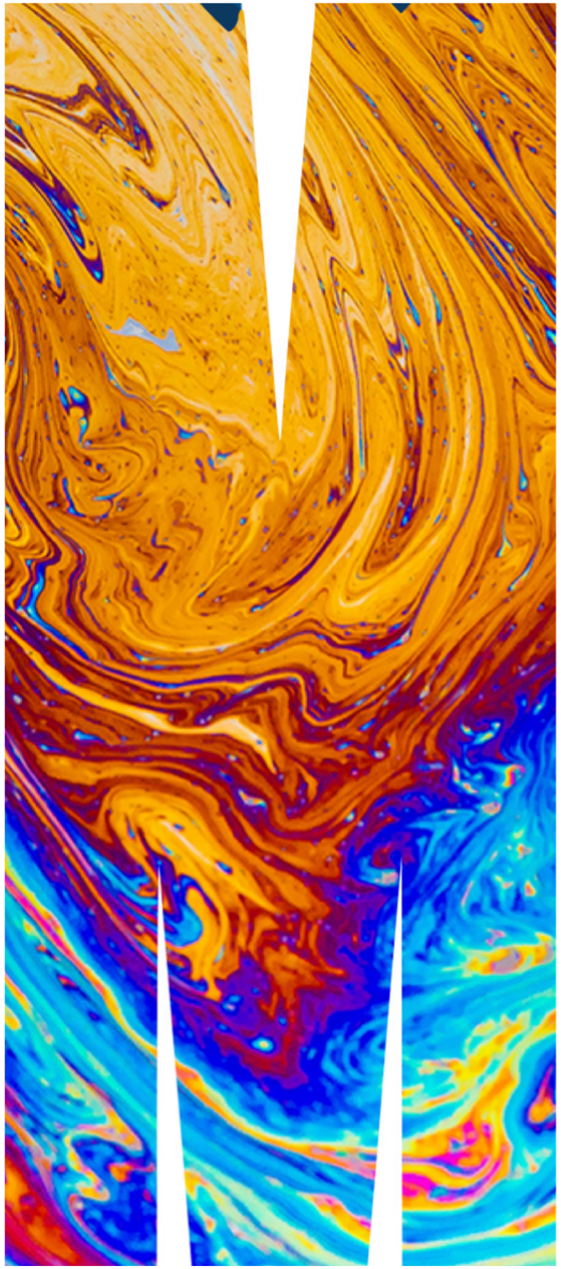
MONASH University

ETC5521: Exploratory Data Analysis

Exploring data having a space and time context



Lecturer: *Di Cook*

✉ ETC5521.Clayton-x@monash.edu

📅 Week 10 - Session 2

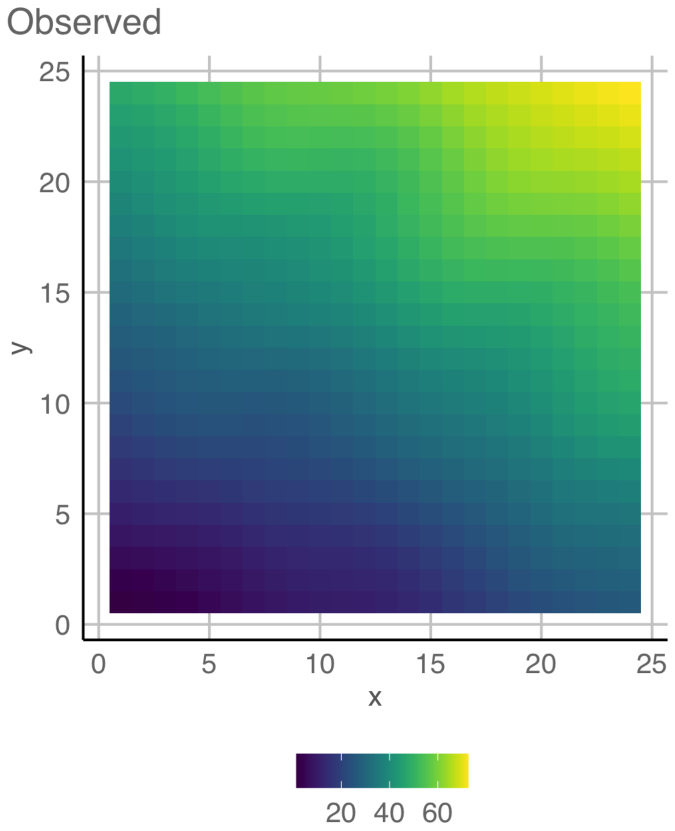
Spatial data

Spatial components

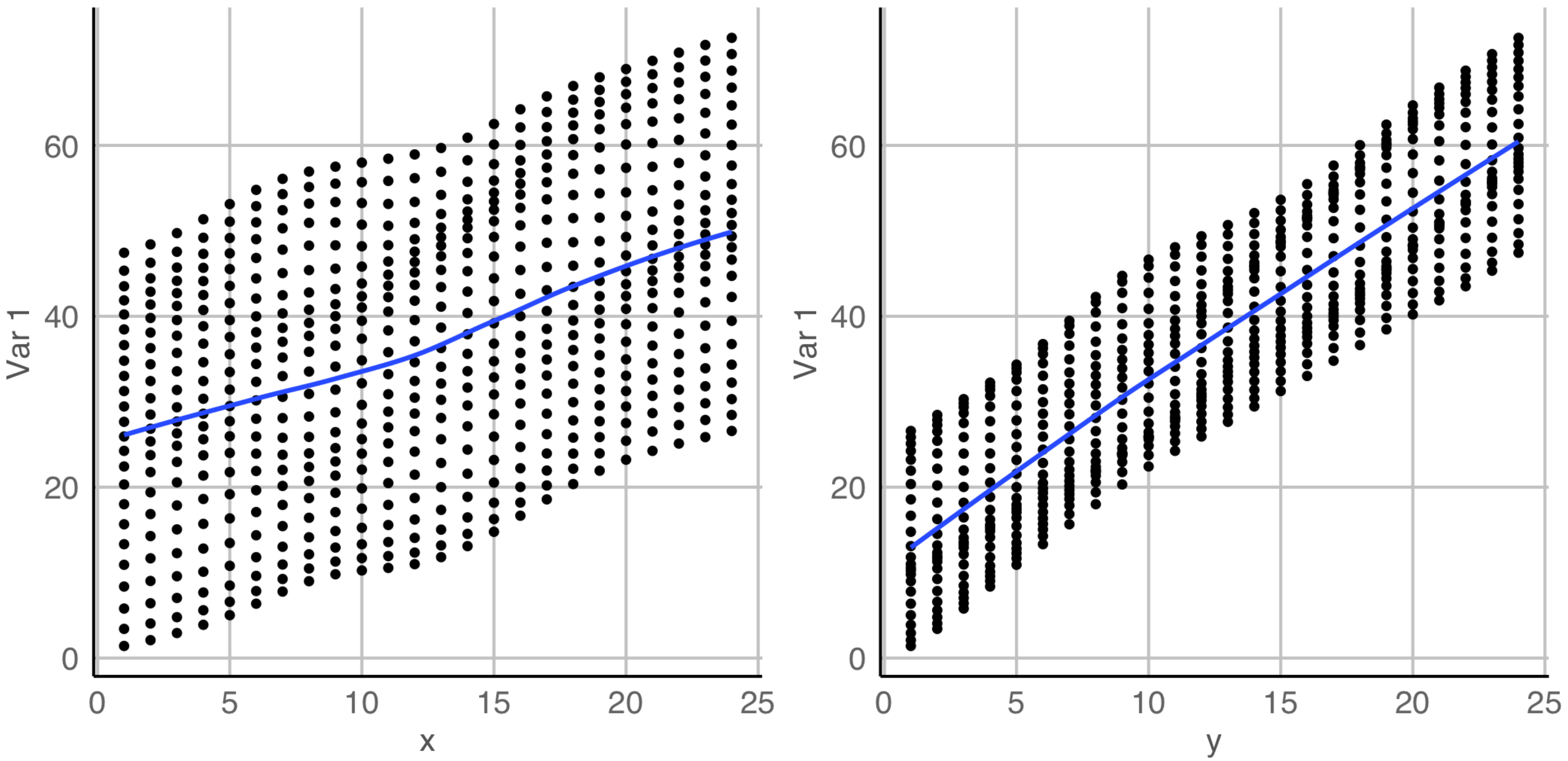
Spatial data can be considered to have both trend and error.

Trend purely on spatial coordinates: expect north-south trend in latitude (position of sun during the year), and possibly east-west in longitude (earth rotation). Trend might be more complicated, localised ecosystems, or related to other factors like elevation.

After trend is removed, the residuals (error) are likely to have spatial dependence: closer sites are likely to have similar values.

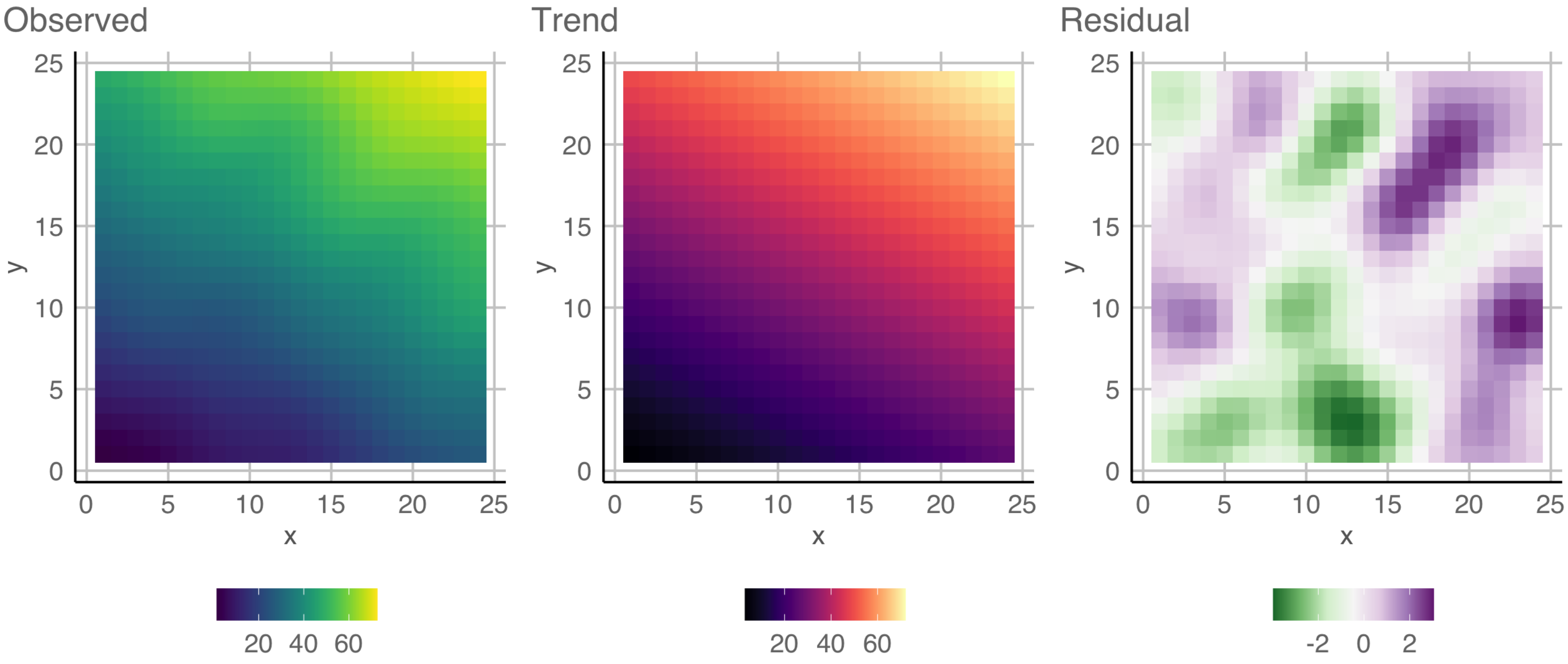


Check trend in longitude and latitude



There is a trend in both directions, but it is stronger in the y (north-south) direction.

Trend + error

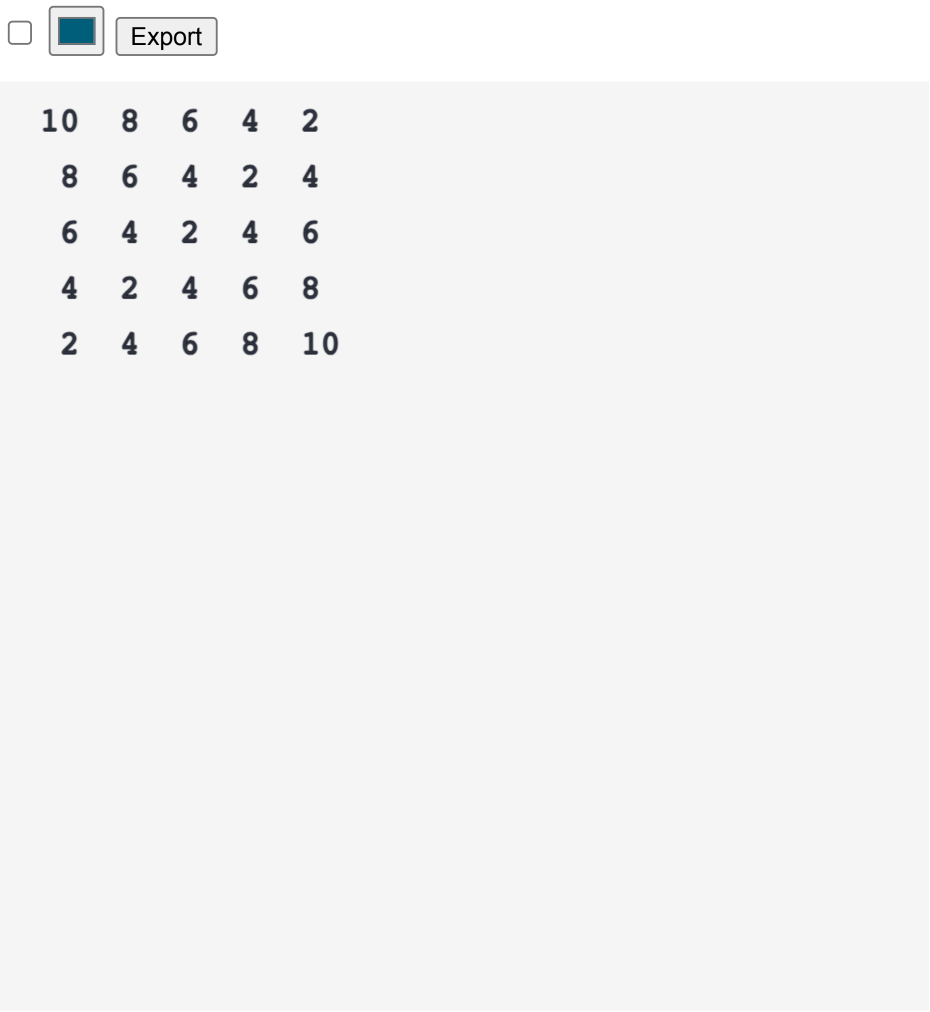


Observed have trend + error. Note the apparent clustering in residuals is strong spatial dependence.

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.

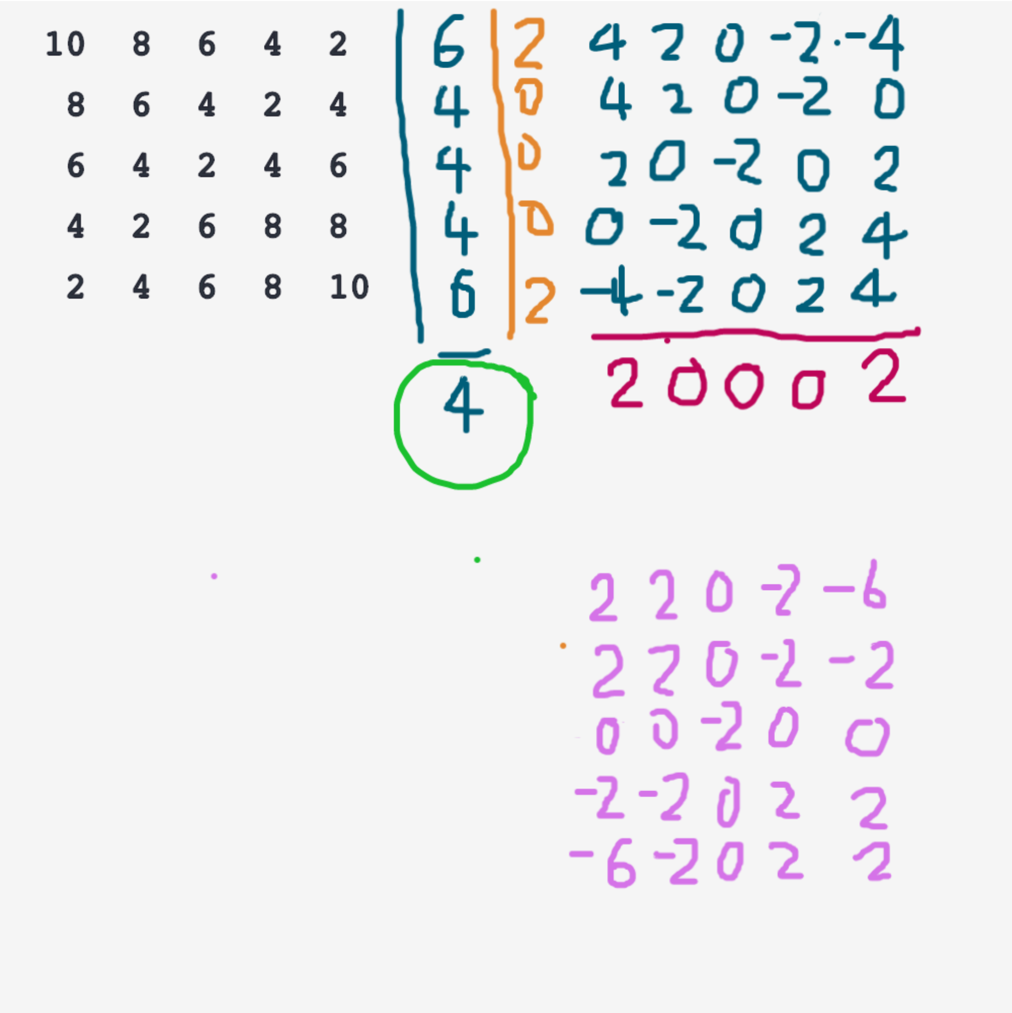
Median polish technique



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

```r
# check calculations
x <- matrix(c(10,  8,  6,  4,  2,
               8,  6,  4,  2,  4,
               6,  4,  2,  4,  6,
               4,  2,  4,  6,  8,
               2,  4,  6,  8,  10),
             nrow=5, byrow=T)
medpolish(x, maxiter = 1)
```

```
## 1: 42

##
## Median Polish Results (Dataset: "x
##
## Overall: 4
```

Median polish technique

```
medpolish(x, maxiter = 5)

## 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
```

Median polish is effectively fitting a model of this form:

overall effect + row effect + column effect

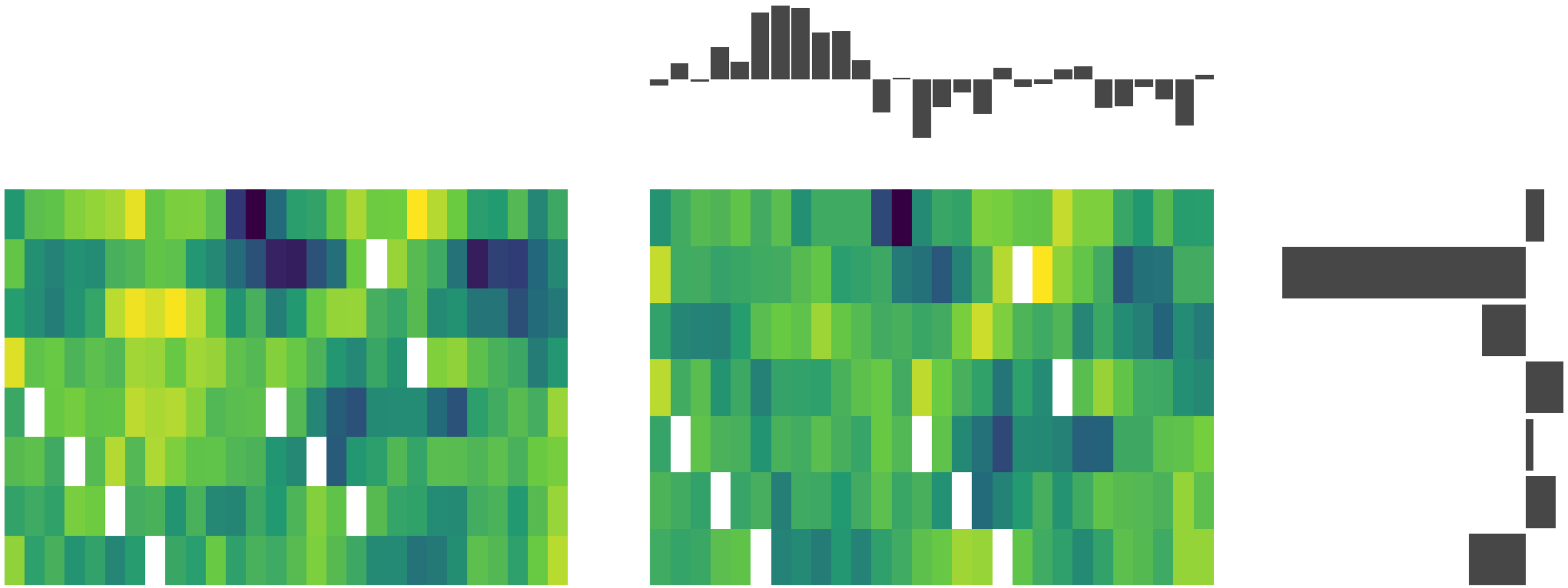
which can be written as:

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

Nice explanation by Manny Gimond

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.

Spatial data needs maps

Maps provide a familiar framework for spatial coordinates.

For data analysis, you want fast to draw maps, not detailed maps.

The important information from maps can be delivered with polygons.

Spatial polygon data

Show 10 entries

Search:

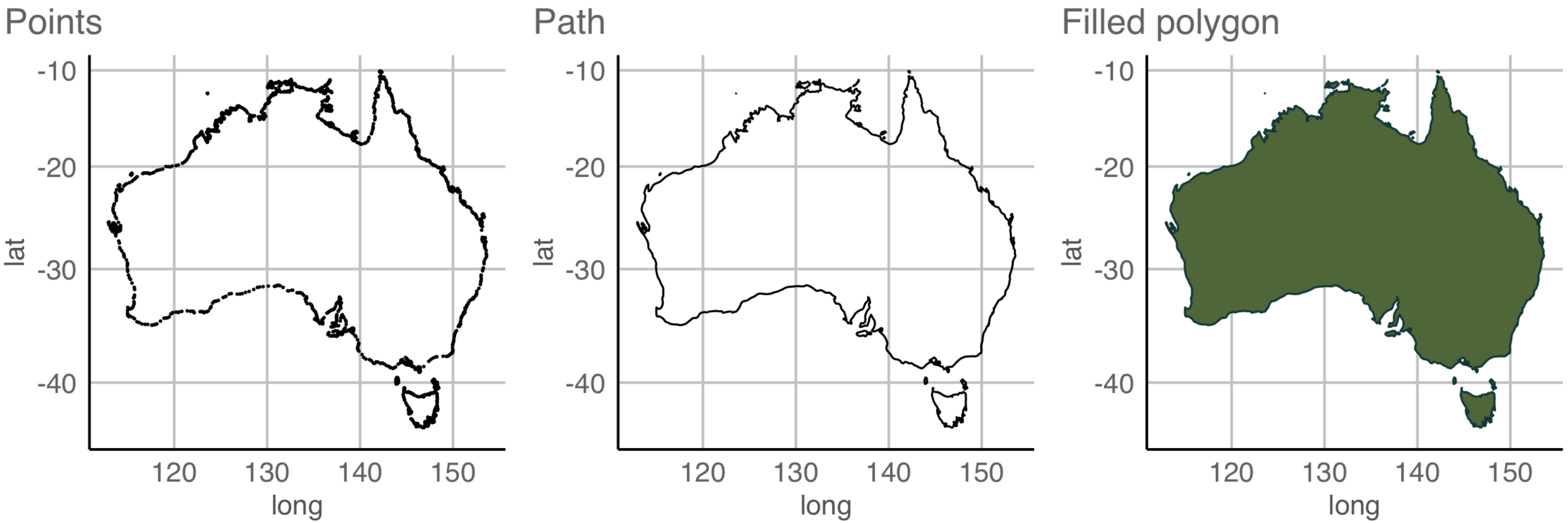
	long	lat	group	order	region	subregion
1	123.5945281982422	-12.42568302154541	133	7115	Australia	Ashmore and Cartier Islands
2	123.5952072143555	-12.43593692779541	133	7116	Australia	Ashmore and Cartier Islands
3	123.5731506347656	-12.43418025970459	133	7117	Australia	Ashmore and Cartier Islands
4	123.5724639892578	-12.42392539978027	133	7118	Australia	Ashmore and Cartier Islands
5	123.5945281982422	-12.42568302154541	133	7119	Australia	Ashmore and Cartier Islands
6	158.8787994384766	-54.70976257324219	139	7267	Australia	Macquarie Island
7	158.84521484375	-54.74921798706055	139	7268	Australia	Macquarie Island
8	158.8359375	-54.70400238037109	139	7269	Australia	Macquarie Island
9	158.89697265625	-54.50605392456055	139	7270	Australia	Macquarie Island
10	158.9588775634766	-54.47236251831055	139	7271	Australia	Macquarie Island

Showing 1 to 10 of 2,579 entries

Previous 1 2 3 4 5 … 258 Next

Spatial polygon data

plot R



Points

Path

Filled polygon

Measured values (variables) associated with a spatial polygon.

sf: Simple spatial polygon objects in R

```
library(sf)
nc <- st_read(system.file("shape/nc.shp", package="sf"))
```

```
## Reading layer `nc' from data source `/Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/library/sf/shape/nc
## Simple feature collection with 100 features and 14 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -84.32385 ymin: 33.88199 xmax: -75.45698 ymax: 36.58965
## Geodetic CRS:  NAD27
```

```
nc %>% slice_head(n=5)
```

```
## Simple feature collection with 5 features and 14 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -81.74107 ymin: 36.07282 xmax: -75.77316 ymax: 36.58965
## Geodetic CRS:  NAD27
##    AREA PERIMETER CNTY_ CNTY_ID        NAME  FIPS FIPSNO CRESS_ID BIR74 SID74 NWBIR74 BIR79 SID79 NWBIR79
## 1 0.114     1.442  1825    1825        Ashe 37009  37009        5  1091     1      10  1364     0      19 MULTIPOLYGO
```

Like the `cubble` object but more strictly a map object. Has a coordinate system (projection), and bounding box. Supports technically accurate distance calculations between coordinates (on a sphere).

sf: Simple spatial polygon objects in R

```
nc_geom <- st_geometry(nc)
nc_geom[[1]] %>% flatten()

## [[1]]
##             [,1]     [,2]
##  [1,] -81.47276 36.23436
##  [2,] -81.54084 36.27251
##  [3,] -81.56198 36.27359
##  [4,] -81.63306 36.34069
##  [5,] -81.74107 36.39178
##  [6,] -81.69828 36.47178
##  [7,] -81.70280 36.51934
##  [8,] -81.67000 36.58965
##  [9,] -81.34530 36.57286
## [10,] -81.34754 36.53791
## [11,] -81.32478 36.51368
## [12,] -81.31332 36.48070
## [13,] -81.26624 36.43721
## [14,] -81.26284 36.40504
## [15,] -81.24069 36.37942
## [16,] -81.23989 36.36536
## [17,] -81.26424 36.35241
```

The geometry contains a list of spatial locations when connected in the right order can be used to draw the spatial polygon.

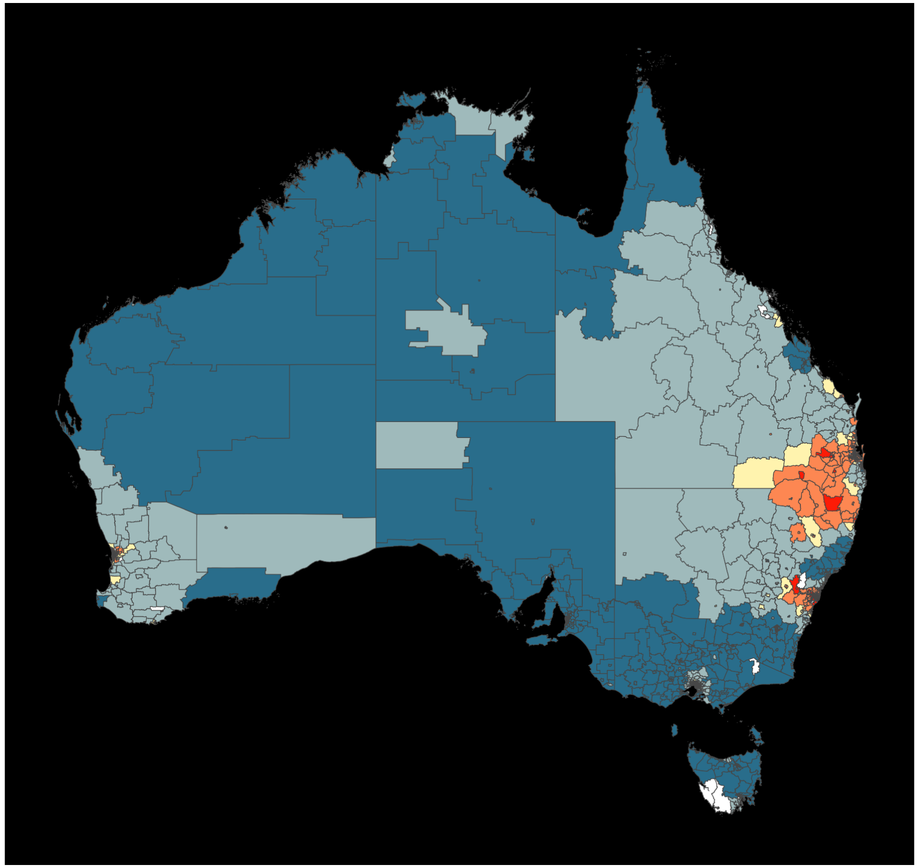
Choropleth maps and cartograms and hexagon tiles

Case study 3 Thyroid cancer in women

plot   learn   R

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.

The purpose is to examine the spatial distribution of a variable.

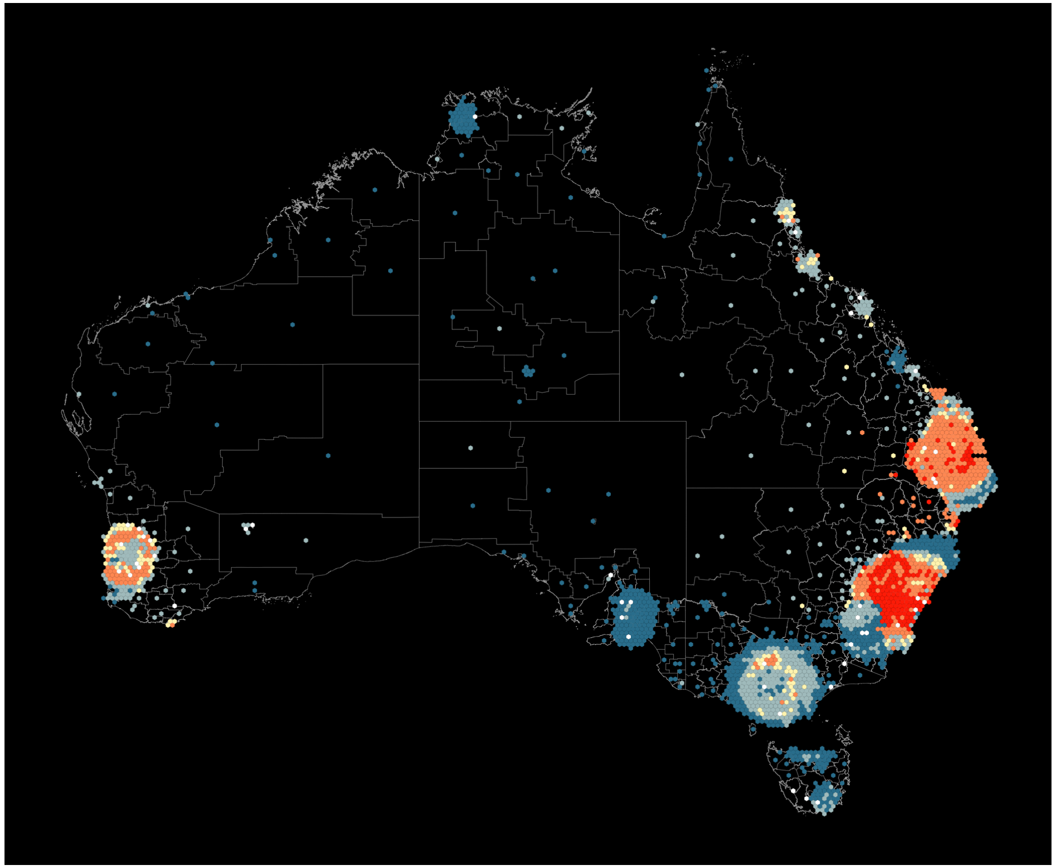


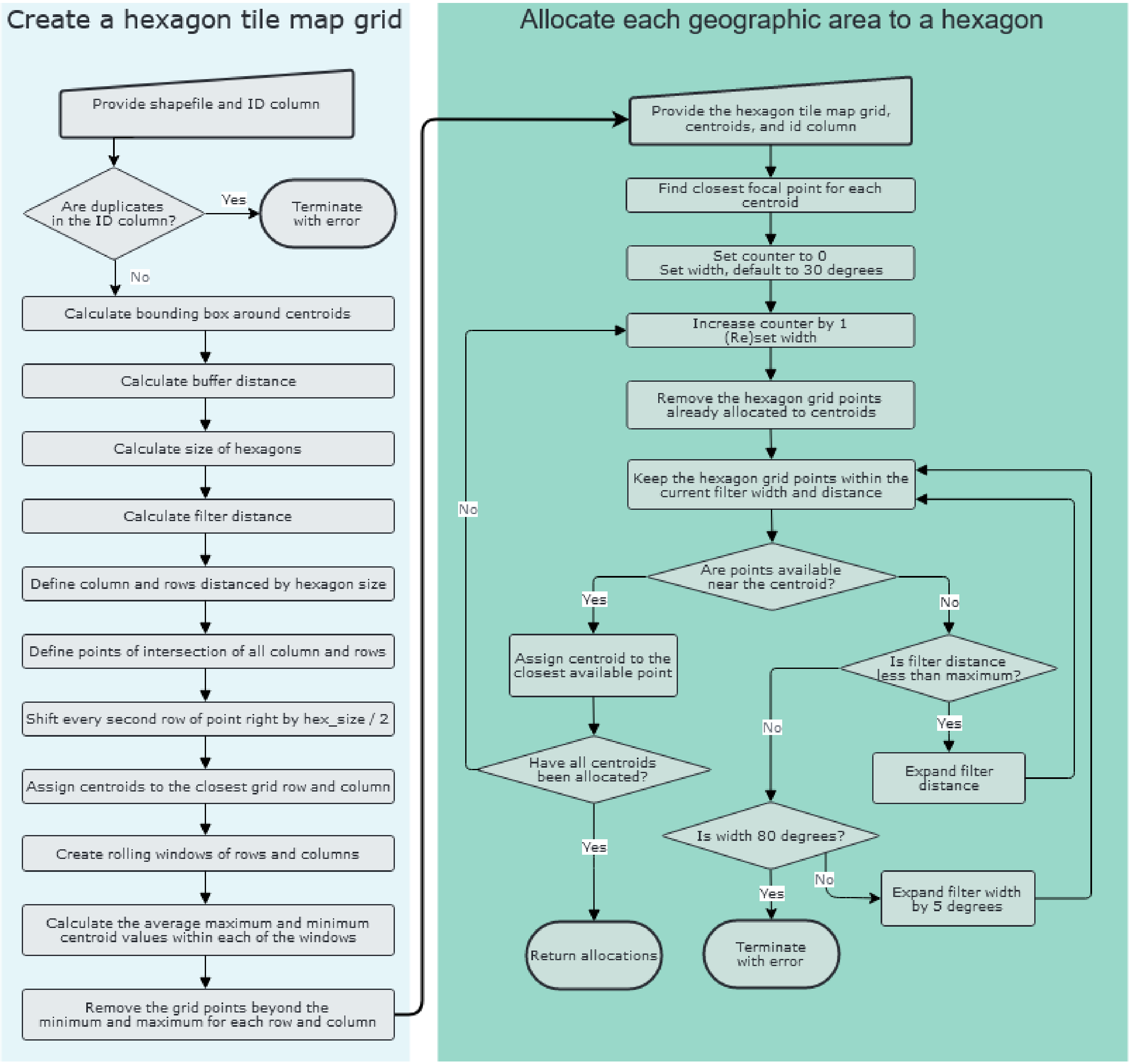
The choropleth map at right shows thyroid cancer incidence for females across Australia, measured at an SA2 level. Red indicates higher incidence.

Case study 3 Thyroid cancer in women

`plot` `learn` `R`

A hexagon tile map represents every spatial polygon with an equal sized hexagon. In dense areas these will be tesselated, but separated hexagons are placed at centroids of the remote spatial regions.





Create a hexagon tile map grid

Provide shapefile and ID column

Are duplicates in the ID column?

Yes

Terminate with error

No

Calculate bounding box around centroids

Calculate buffer distance

Calculate size of hexagons

Calculate filter distance

Define column and rows distanced by hexagon size

Define points of intersection of all column and rows

Shift every second row of point right by hex_size / 2

Assign centroids to the closest grid row and column

Create rolling windows of rows and columns

Calculate the average maximum and minimum centroid values within each of the windows

Remove the grid points beyond the minimum and maximum for each row and column

Allocate each geographic area to a hexagon

Provide the hexagon tile map grid, centroids, and id column

Find closest focal point for each centroid

Set counter to 0
Set width, default to 30 degrees

Increase counter by 1
(Re)set width

Remove the hexagon grid points already allocated to centroids

Keep the hexagon grid points within the current filter width and distance

Are points available near the centroid?

Yes

Assign centroid to the closest available point

Have all centroids been allocated?

No

Yes

Return allocations

No

Is filter distance less than maximum?

Yes

Expand filter distance

No

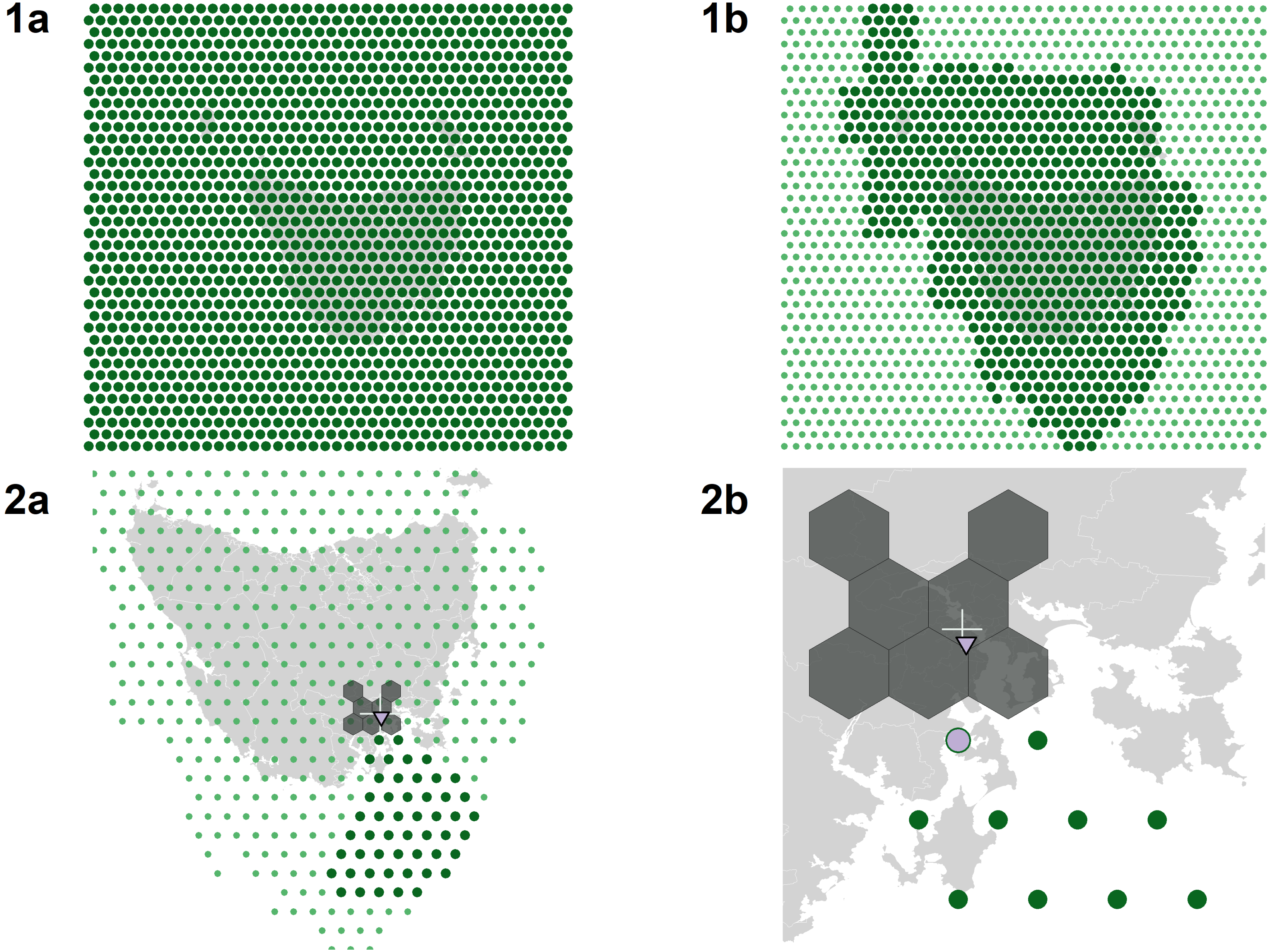
Is width 80 degrees?

Yes

Terminate with error

No

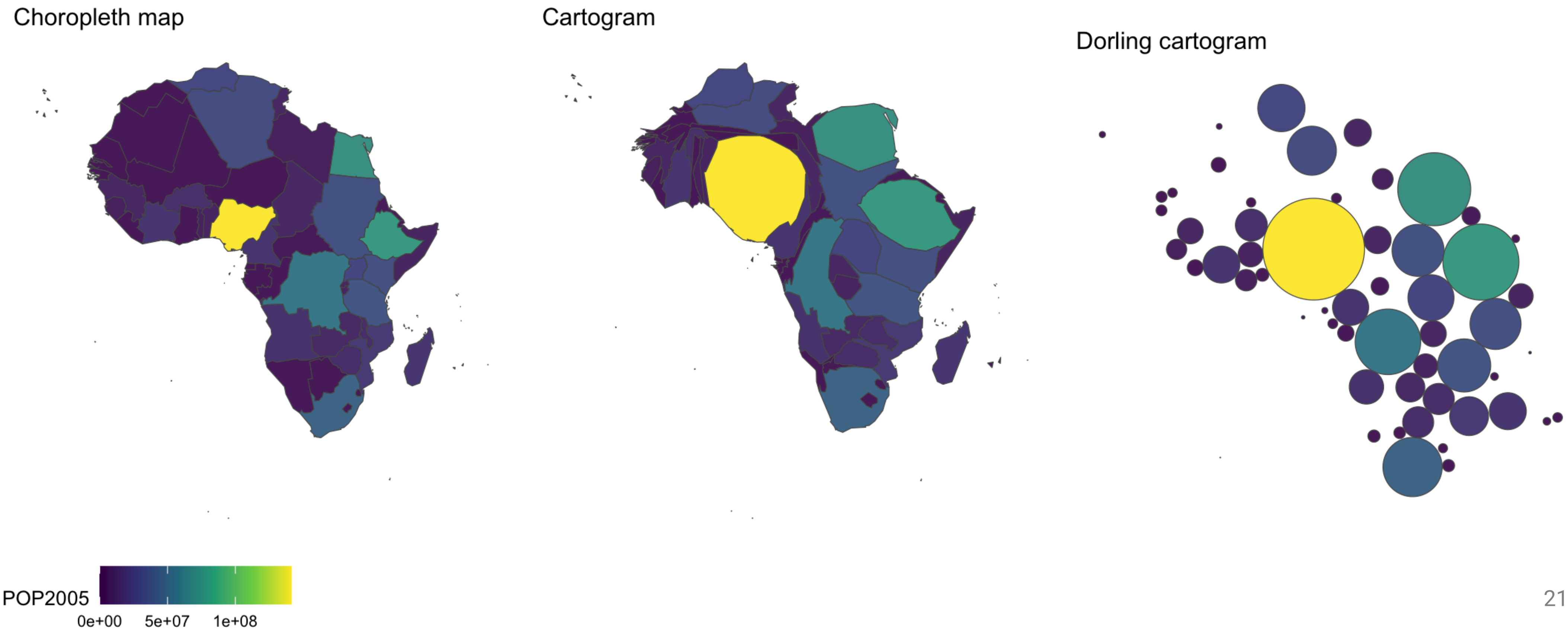
Expand filter width by 5 degrees


1a
1b
2a
2b

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



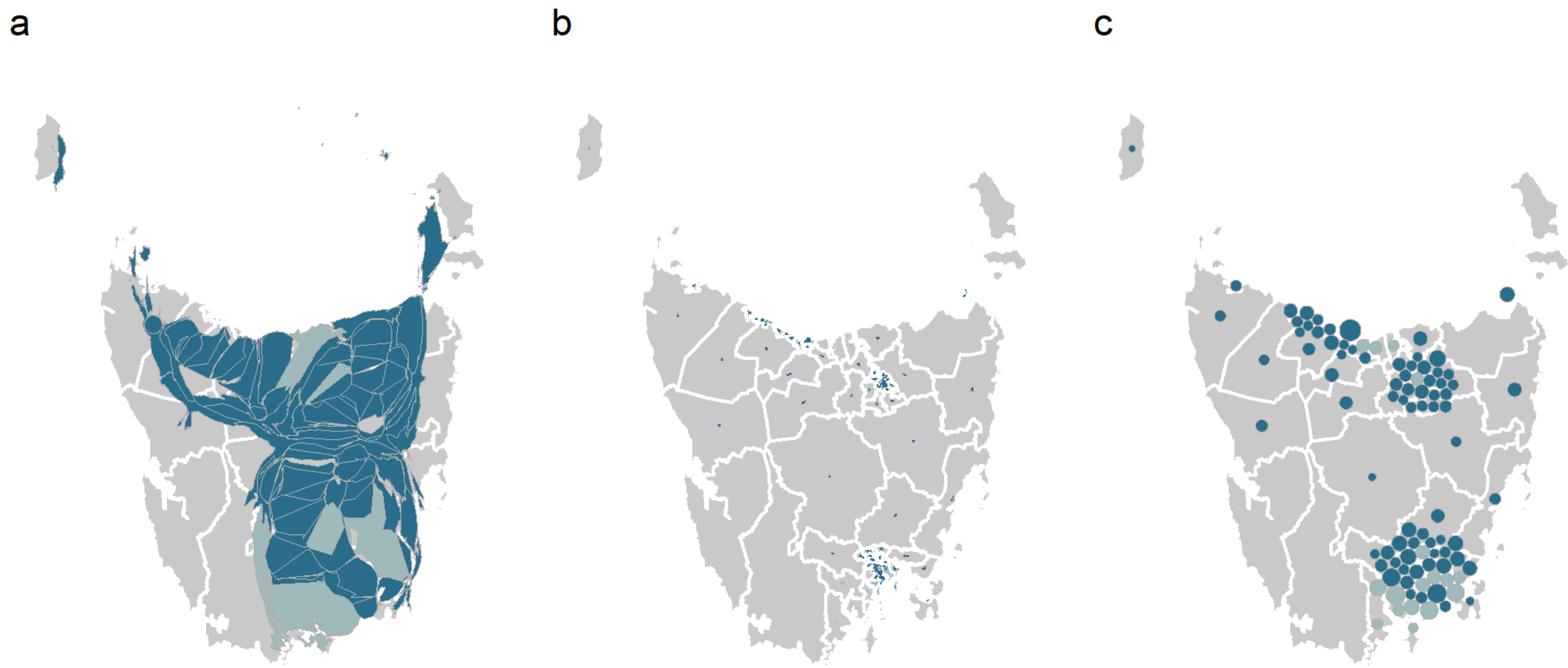
Show 10 entries

Search:

	FIPS	ISO2	ISO3	UN	NAME	AREA	POP2005	REGION	SUBREGION	LON	LAT	geometry
DZA	AG	DZ	DZA	12	Algeria	238174	32854159	2	15	2.632	28.163	[object Object]
AGO	AO	AO	AGO	24	Angola	124670	16095214	2	17	17.544	-12.296	[object Object]
BEN	BN	BJ	BEN	204	Benin	11062	8490301	2	11	2.469	10.541	[object Object]
COG	CF	CG	COG	178	Congo	34150	3609851	2	17	15.986	-0.055	[object Object]
COD	CG	CD	COD	180	Democratic Republic of the Congo	226705	58740547	2	17	23.654	-2.876	[object Object]
BDI	BY	BI	BDI	108	Burundi	2568	7858791	2	14	29.887	-3.356	[object Object]
CMR	CM	CM	CMR	120	Cameroon	46540	17795149	2	17	12.277	5.133	[object Object]
TCD	CD	TD	TCD	148	Chad	125920	10145609	2	17	18.665	15.361	[object Object]
COM	CN	KM	COM	174	Comoros	223	797902	2	14	43.337	-11.758	[object Object]
CAF	CT	CF	CAF	140	Central African Republic	62298	4191429	2	17	20.483	6.571	[object Object]

Showing 1 to 10 of 57 entries

Previous 1 2 3 4 5 6 Next



Three different cartogram displays for Tasmania: (a) contiguous cartogram, (b) non-contiguous cartogram and (c) Dorling cartogram.

The cartogram algorithm can dramatically alter the geography, so that it is no longer recognisable. In the case of the whole of Australia, it simply does not converge.

Your turn! Point your browser to Michael Gastner's cartogram web site:

https://go-cart.io/cartogram

Resources and Acknowledgement

- Wickham et al (2012) Glyph-maps for Visually Exploring Temporal Patterns in Climate Data and Models
- sf: Simple Features for R
- Hexmaps with sugarbag and documentation
- Making cartograms in R
- Gastner et al (2018) Fast flow-based algorithm for creating density-equalizing map projections
- Median polish on two way tables from Tukey, J. W. (1977). Exploratory Data Analysis, Reading Massachusetts: Addison-Wesley, see Manny Gimond's explanation.

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 10 - Session 2