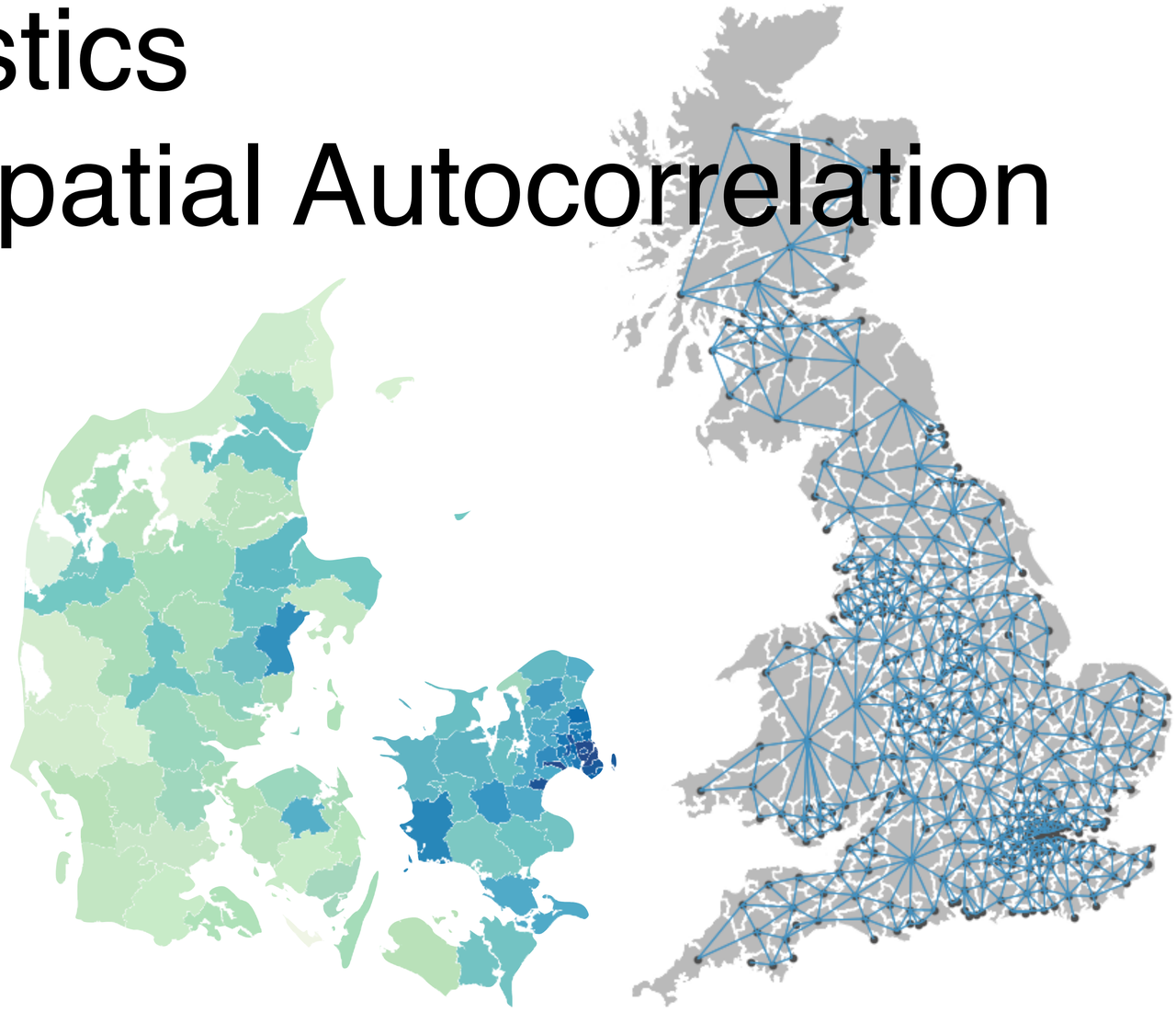
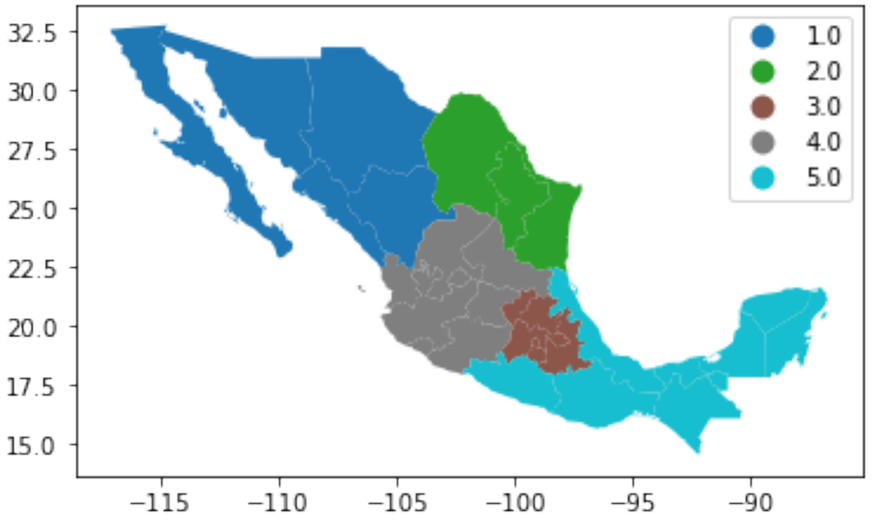
Part 2: Spatial Statistics
Choropleth Maps, Spatial Autocorrelation

Michael Szell
Jul 17, 2023



In this session you will learn about visual and statistical basics

Choropleths: Theory & Python

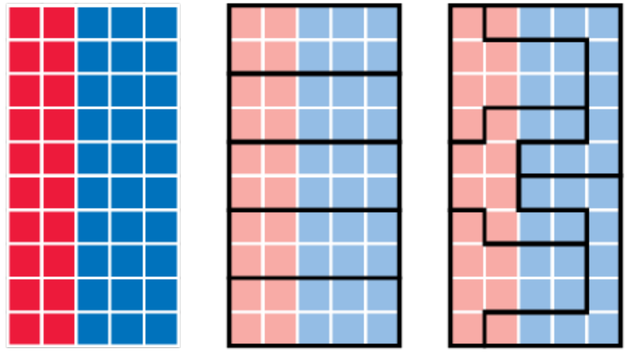


Global and local measures of spatial autocorrelation

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j \; ; \; m_2 = \frac{\sum_i z_i^2}{n}$$

Pitfalls and biases



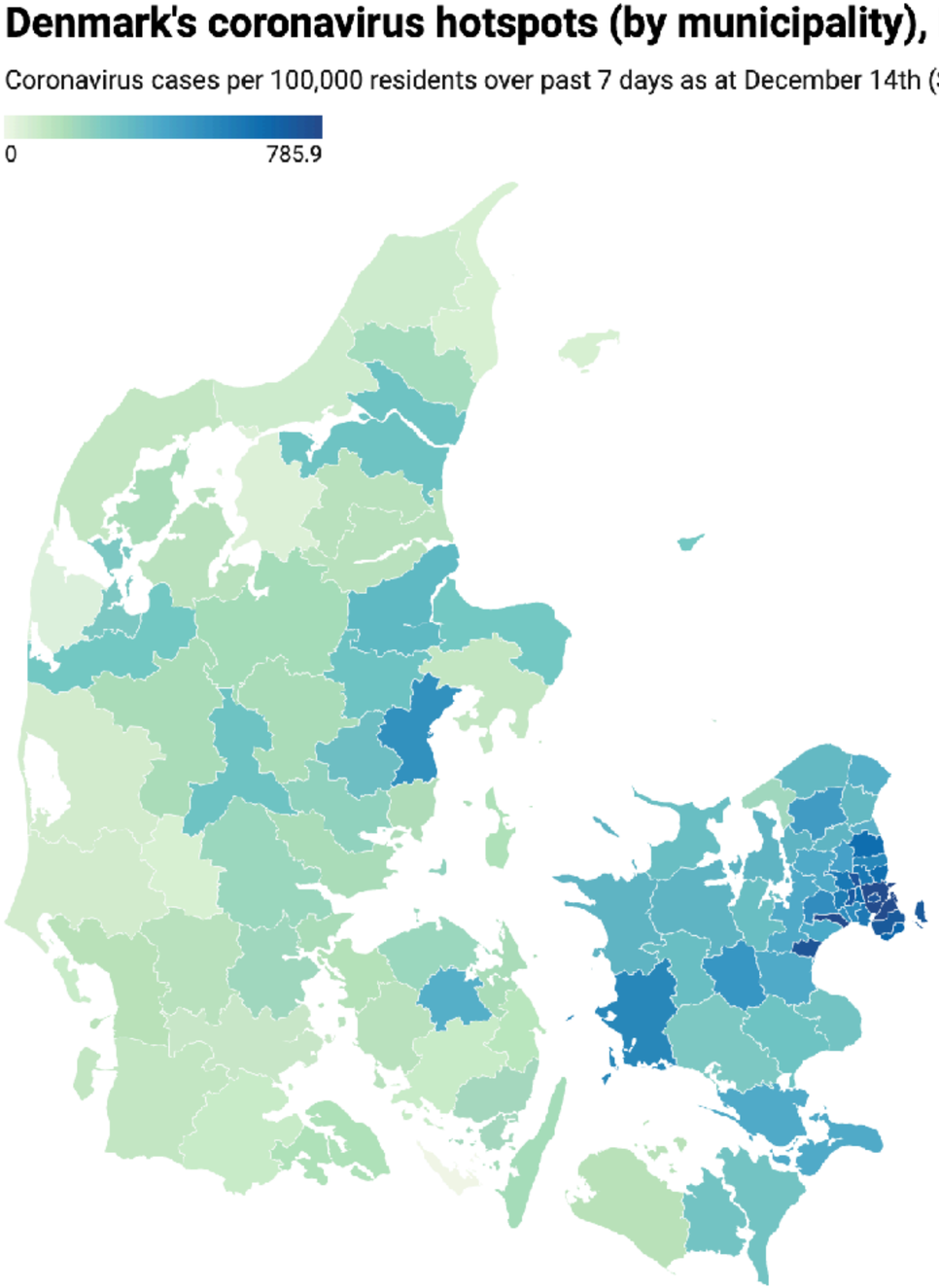
Aggregation = Combining multiple objects into a single one

Aggregation = Combining multiple objects into a single one

GPS coordinate → Zip Code → City → Country

Advantages: Data reduction, easier to process, high-level view, smaller statistical fluctuations

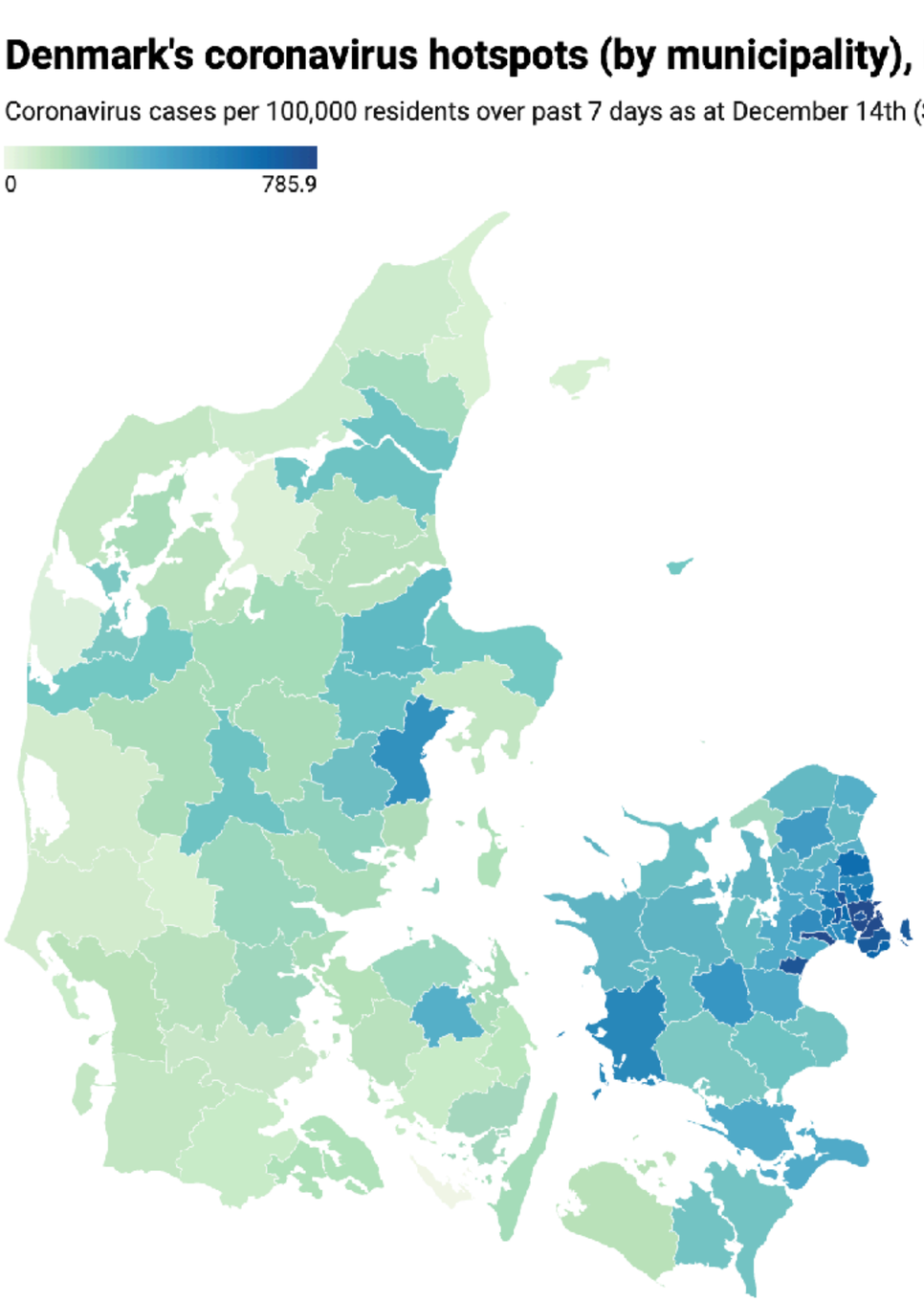
Disadvantages: Loss of details, introducing biases



‘Choro’ = region
‘pleth’ = multiple

3 choices:

https://www.thelocal.dk/20201028/maps-where-are-denmarks-coronavirus-hotspots/



‘Choro’ = region
‘pleth’ = multiple

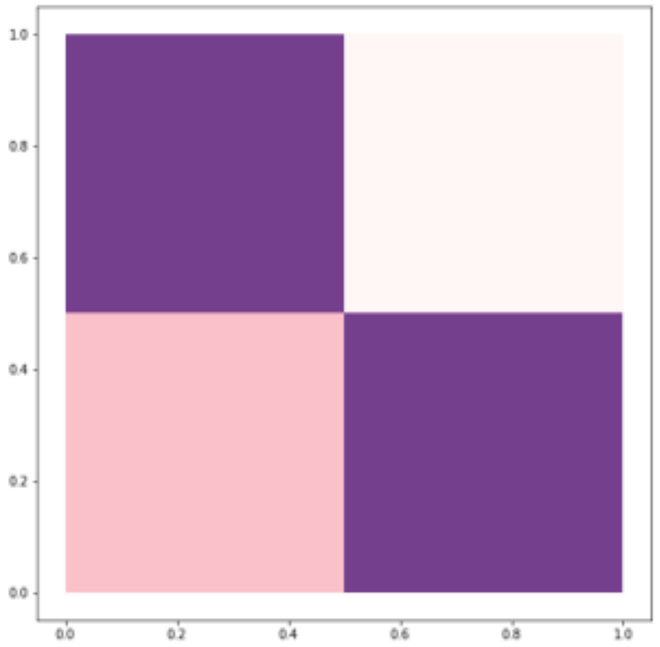
3 choices:

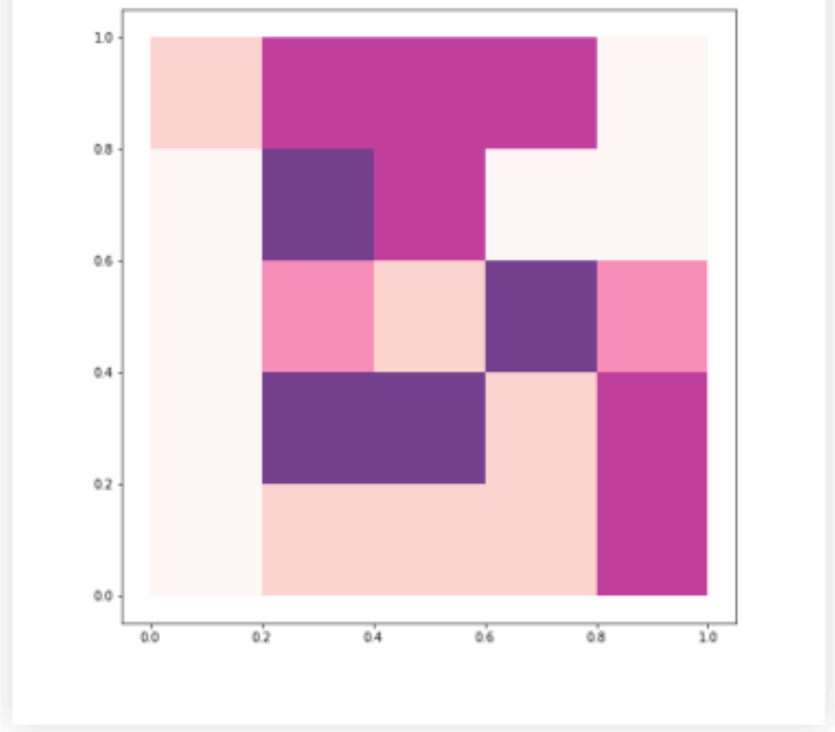
1) Spatial units
2) Colors
3) Classes

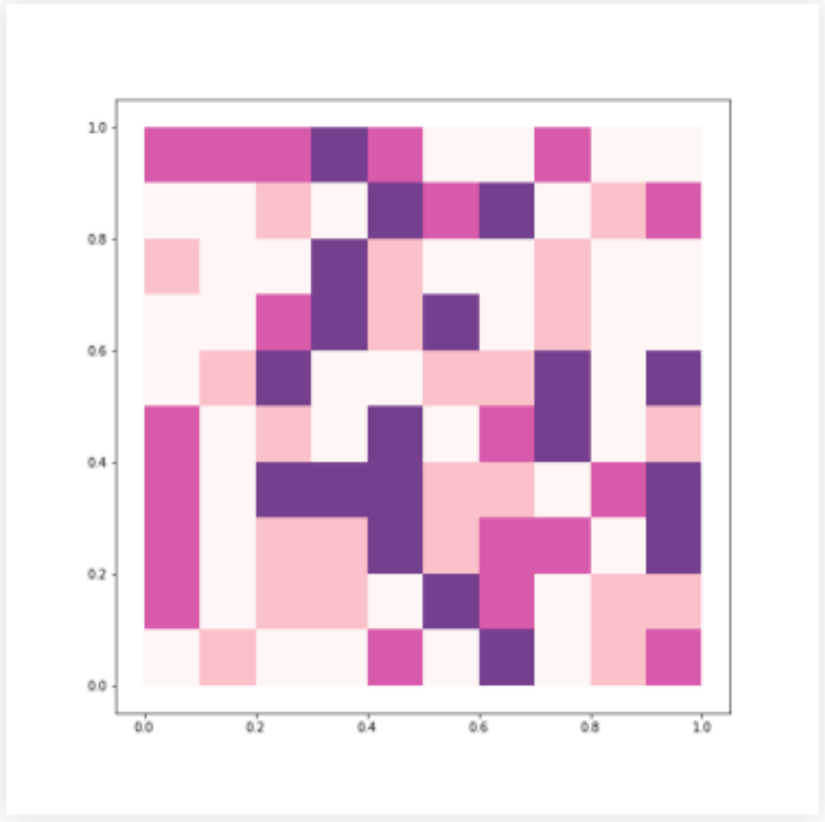
https://www.thelocal.dk/20201028/maps-where-are-denmarks-coronavirus-hotspots/

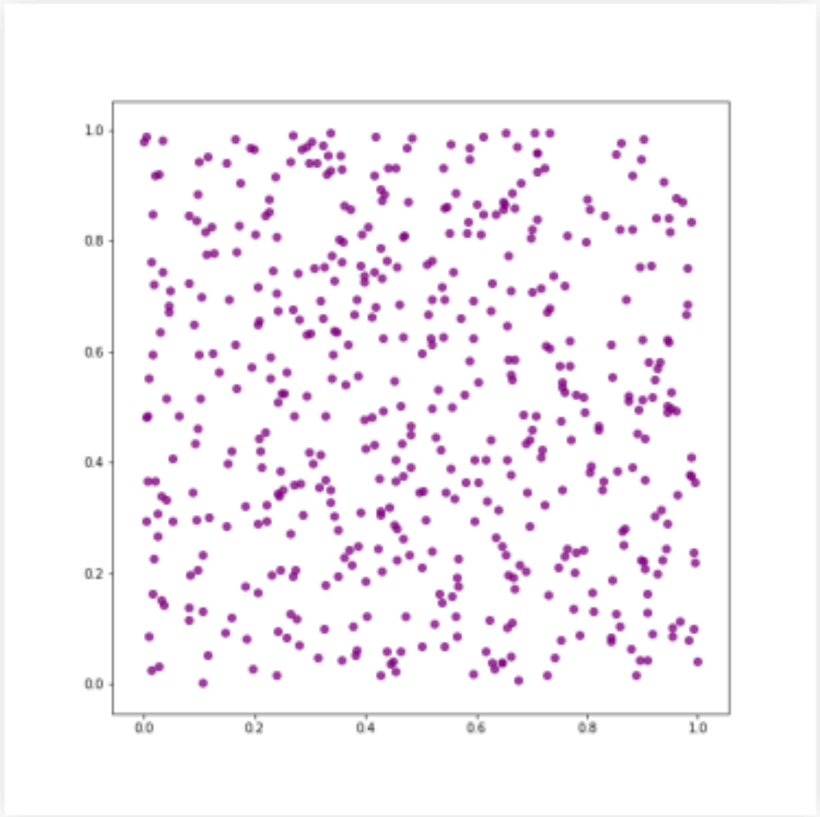
1) Spatial units

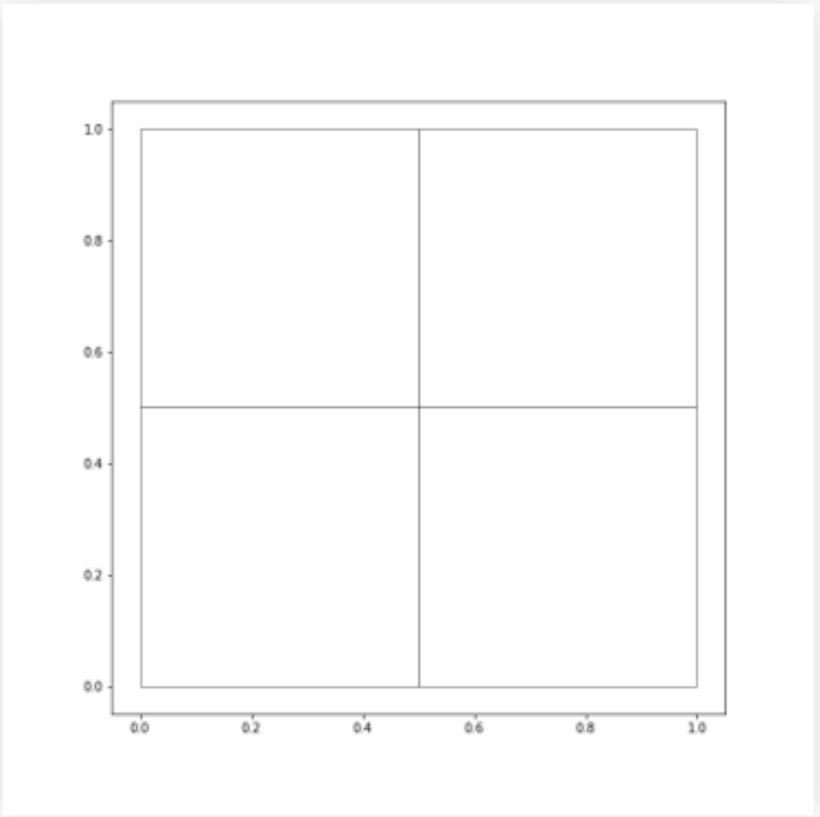
Describe the population distribution on the map

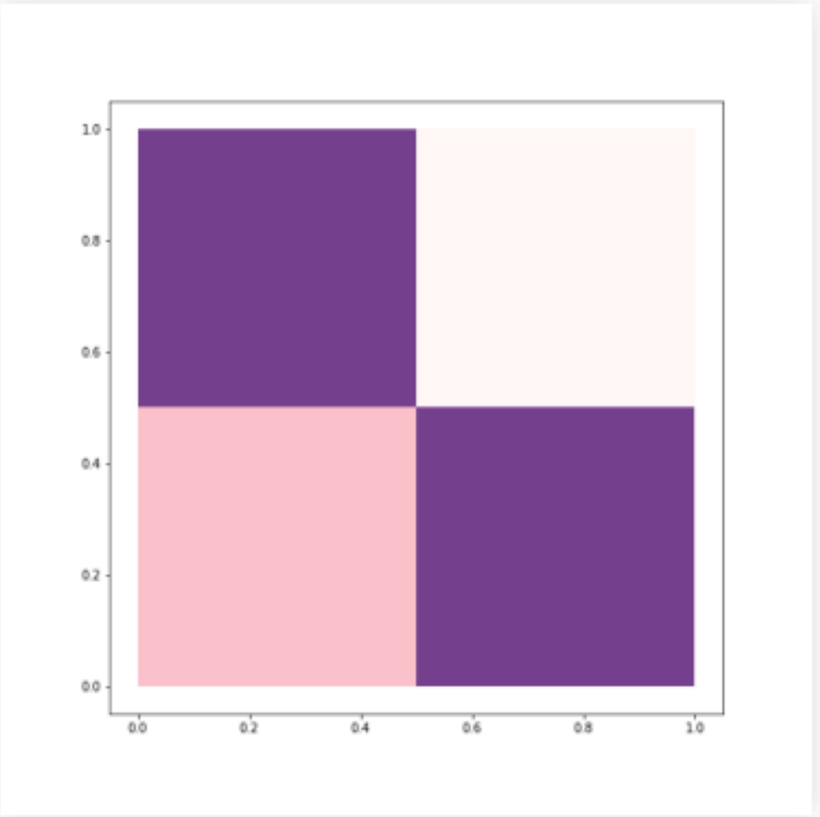

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

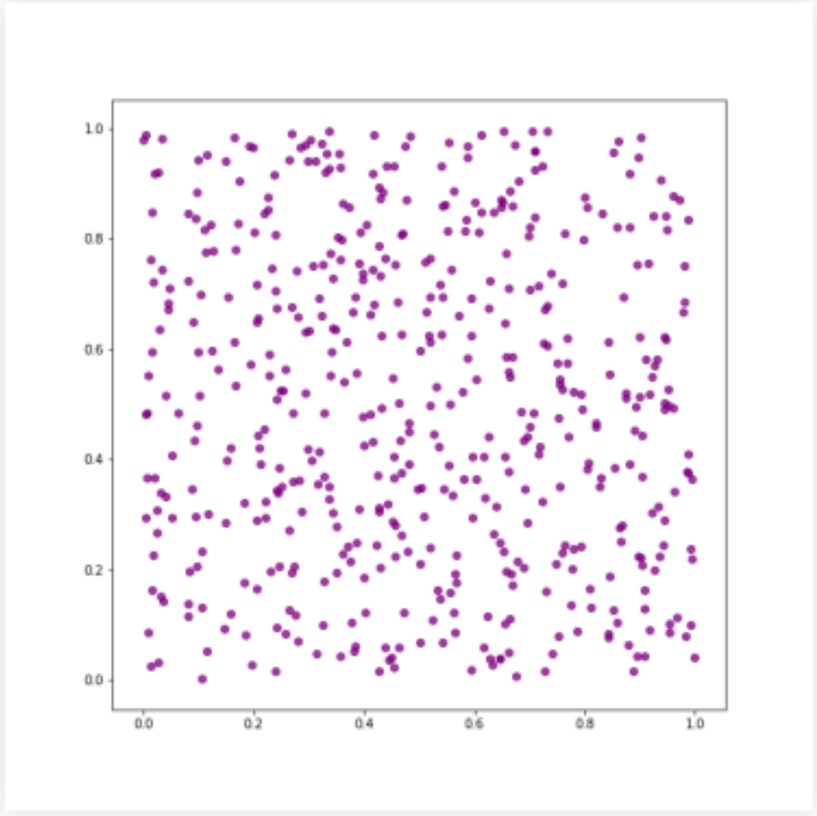

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

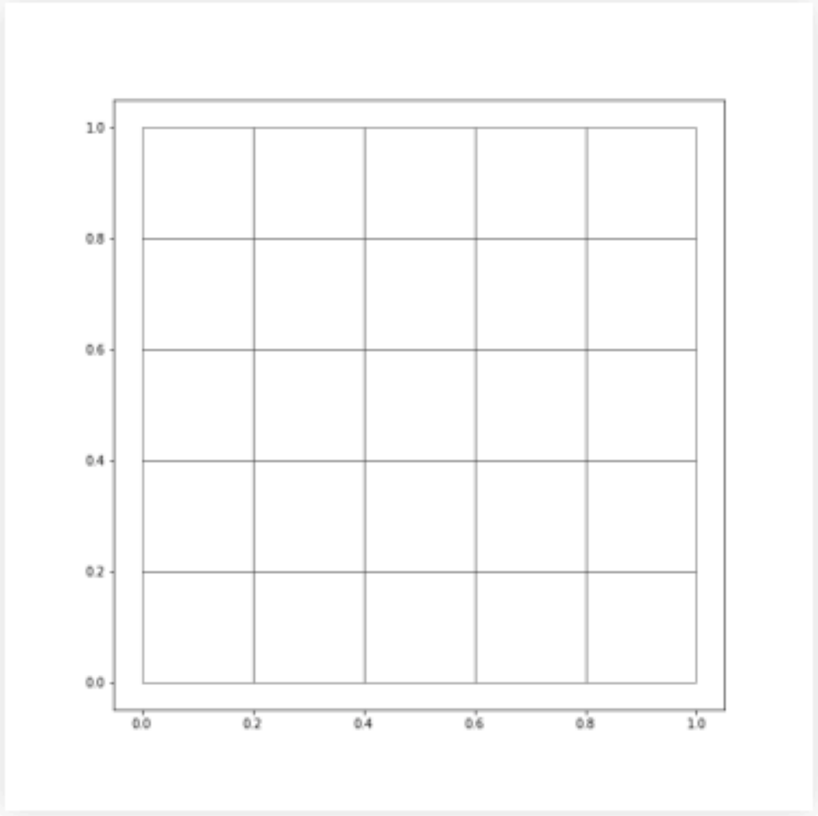


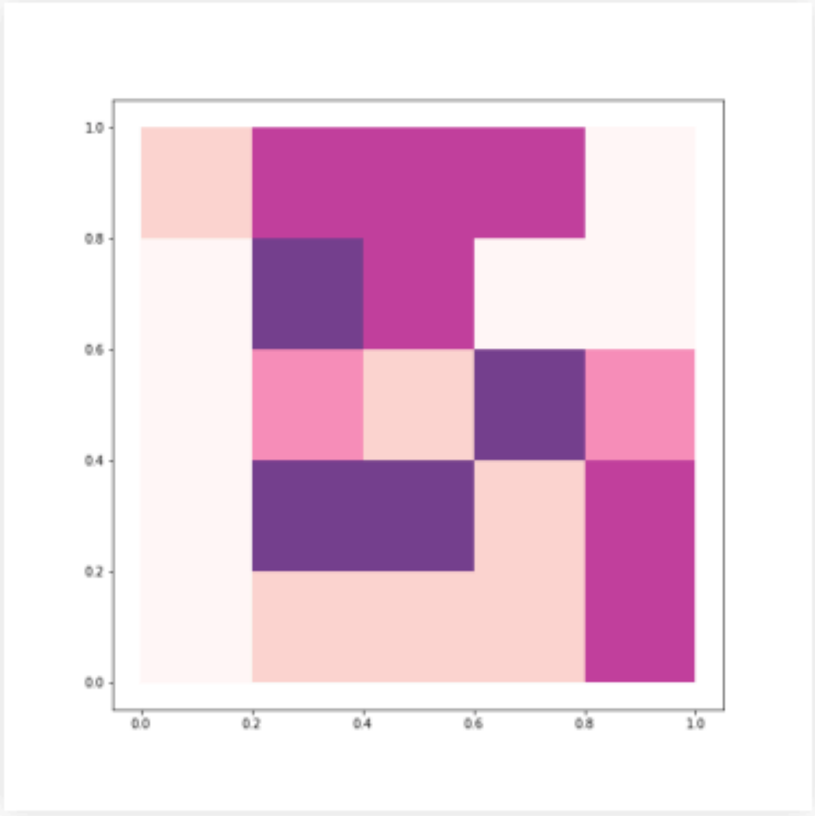




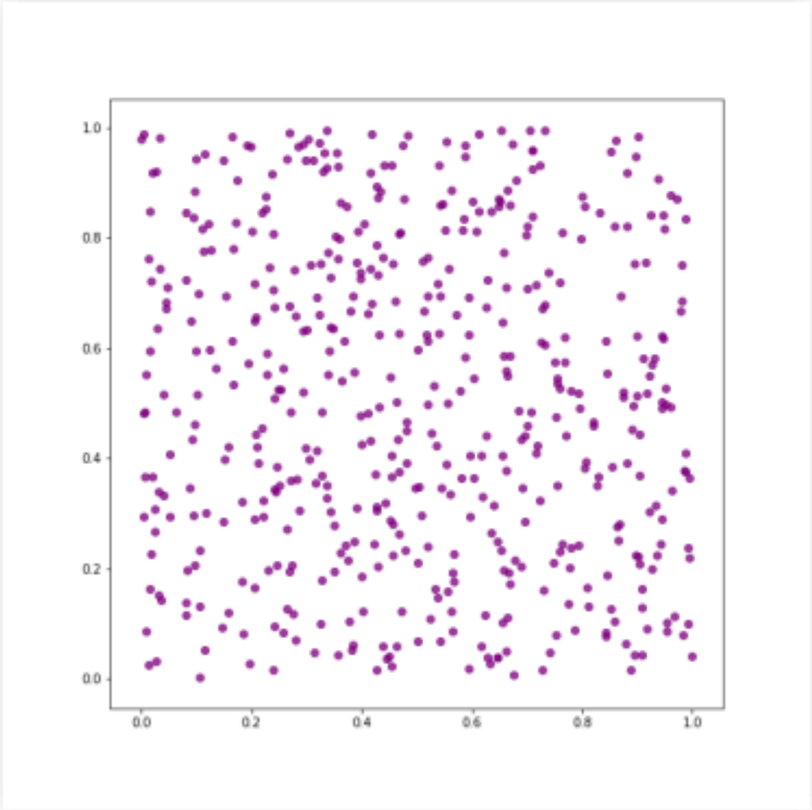
https://gist.github.com/darribas/8b5a7b93d4085223f1c5

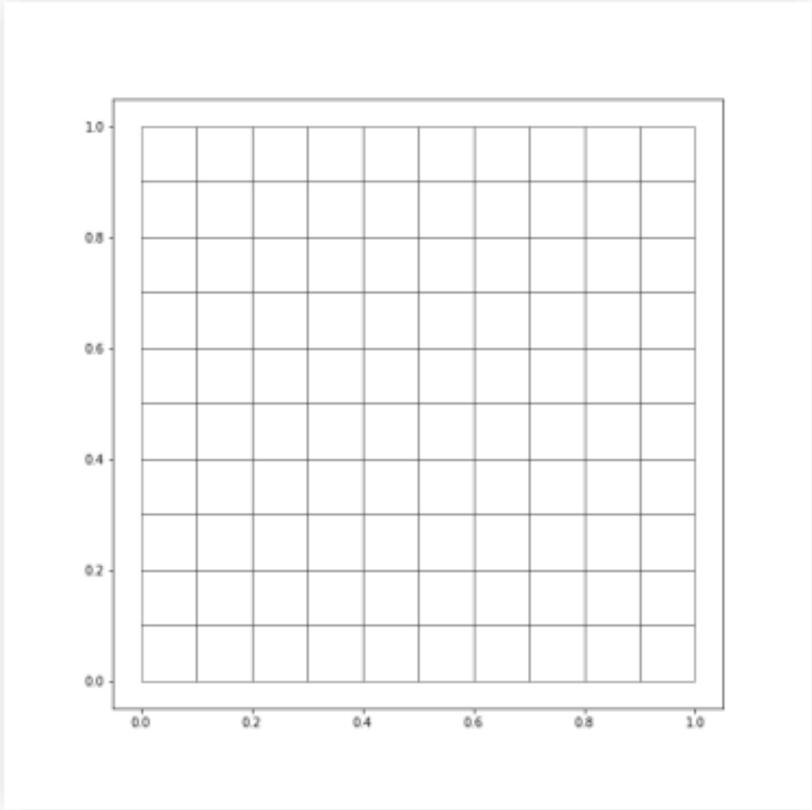


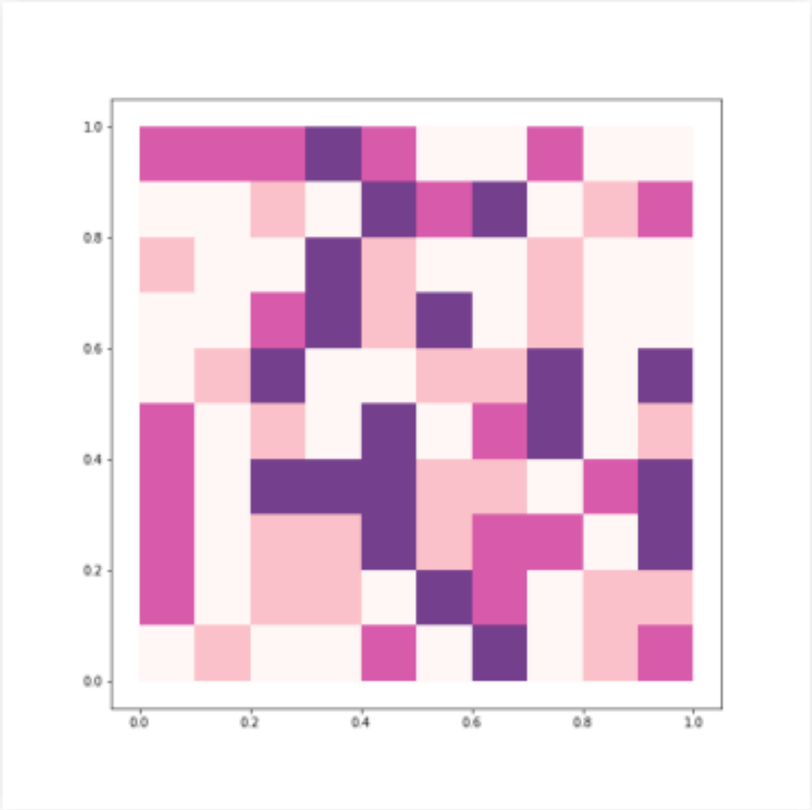




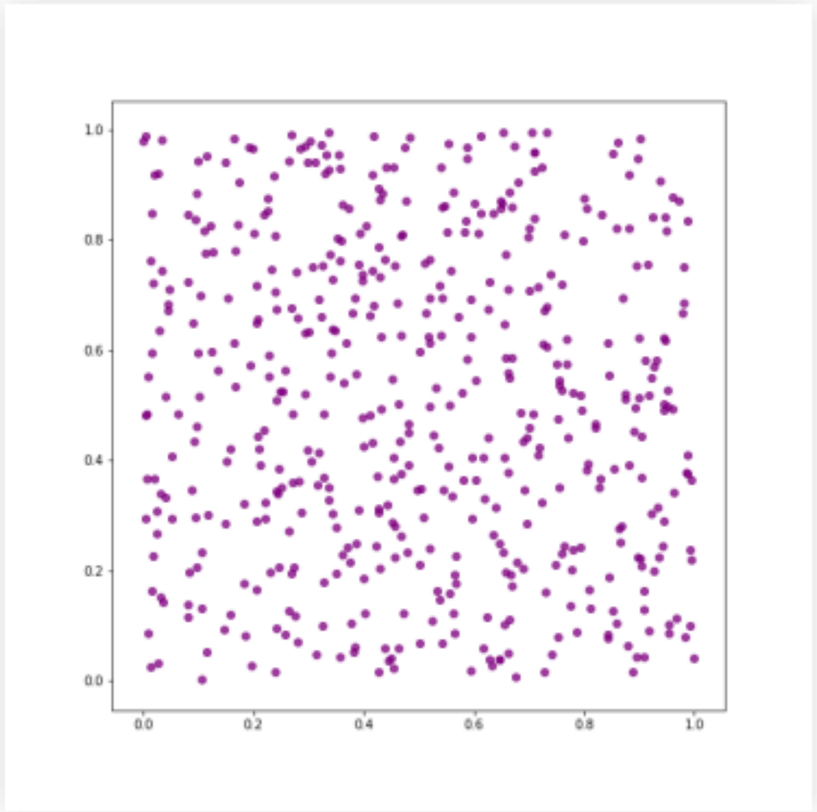
https://gist.github.com/darribas/8b5a7b93d4085223f1c5

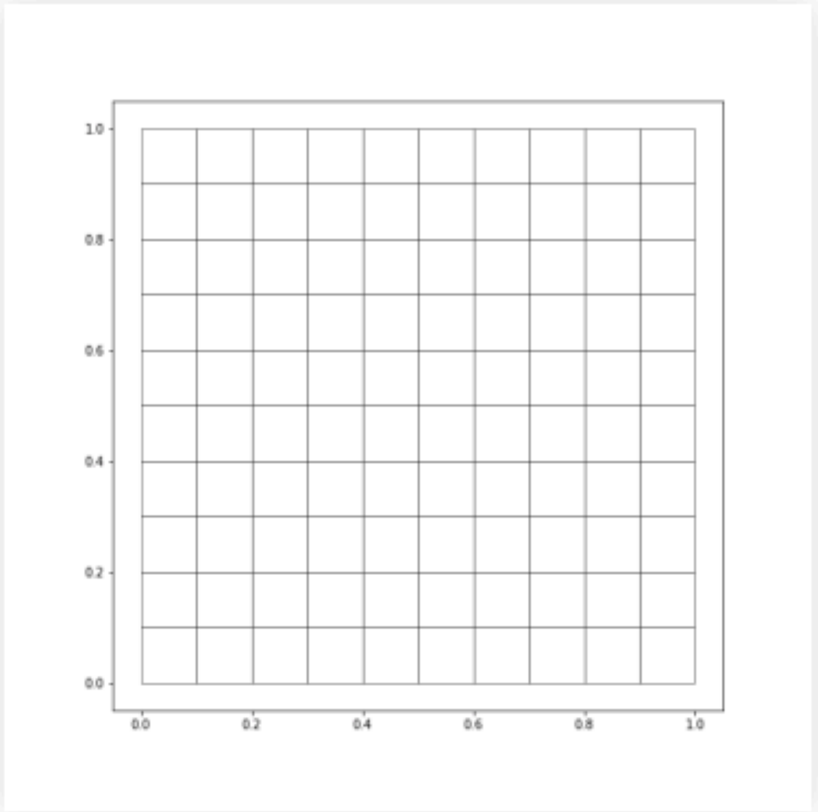


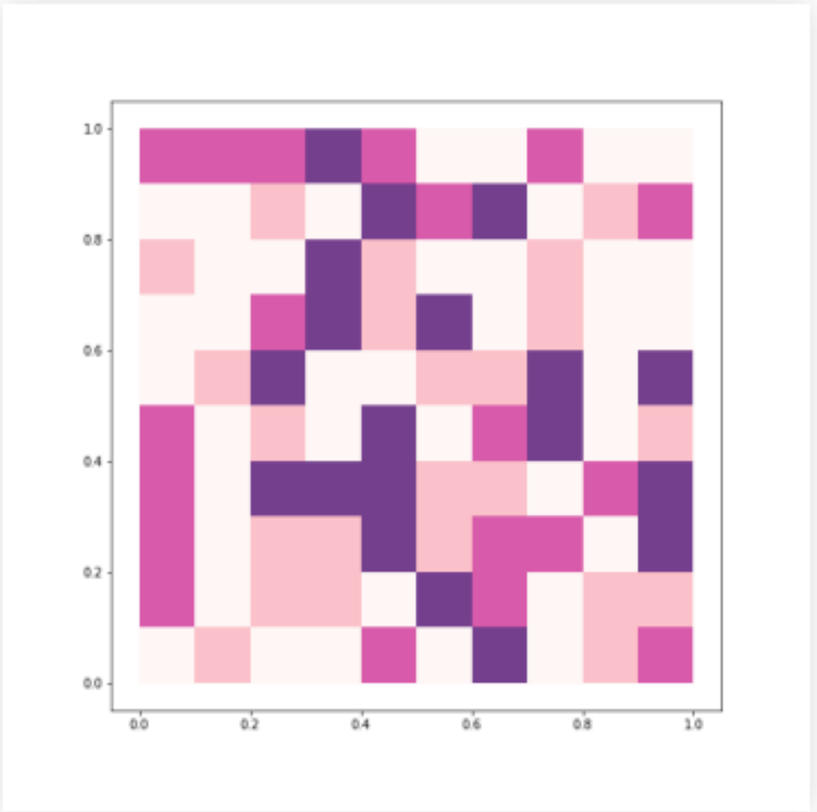




https://gist.github.com/darribas/8b5a7b93d4085223f1c5







Same data, Same colors, Same classification,

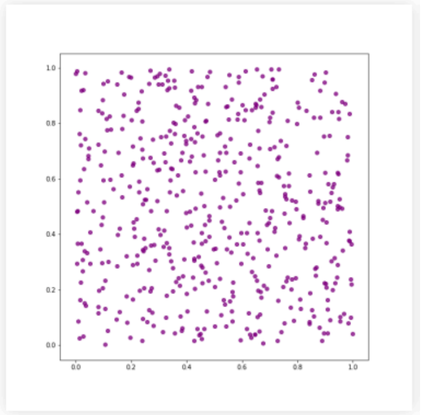
Different spatial units

https://gist.github.com/darribas/8b5a7b93d4085223f1c5

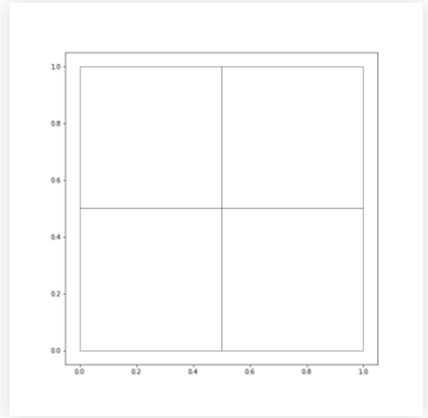
A common source of bias in spatial aggregation: MAUP

The MAUP (Modifiable Areal Unit Problem) is a scale and delineation mismatch between:

Underlying entities $\longleftrightarrow$ Unit of measurement



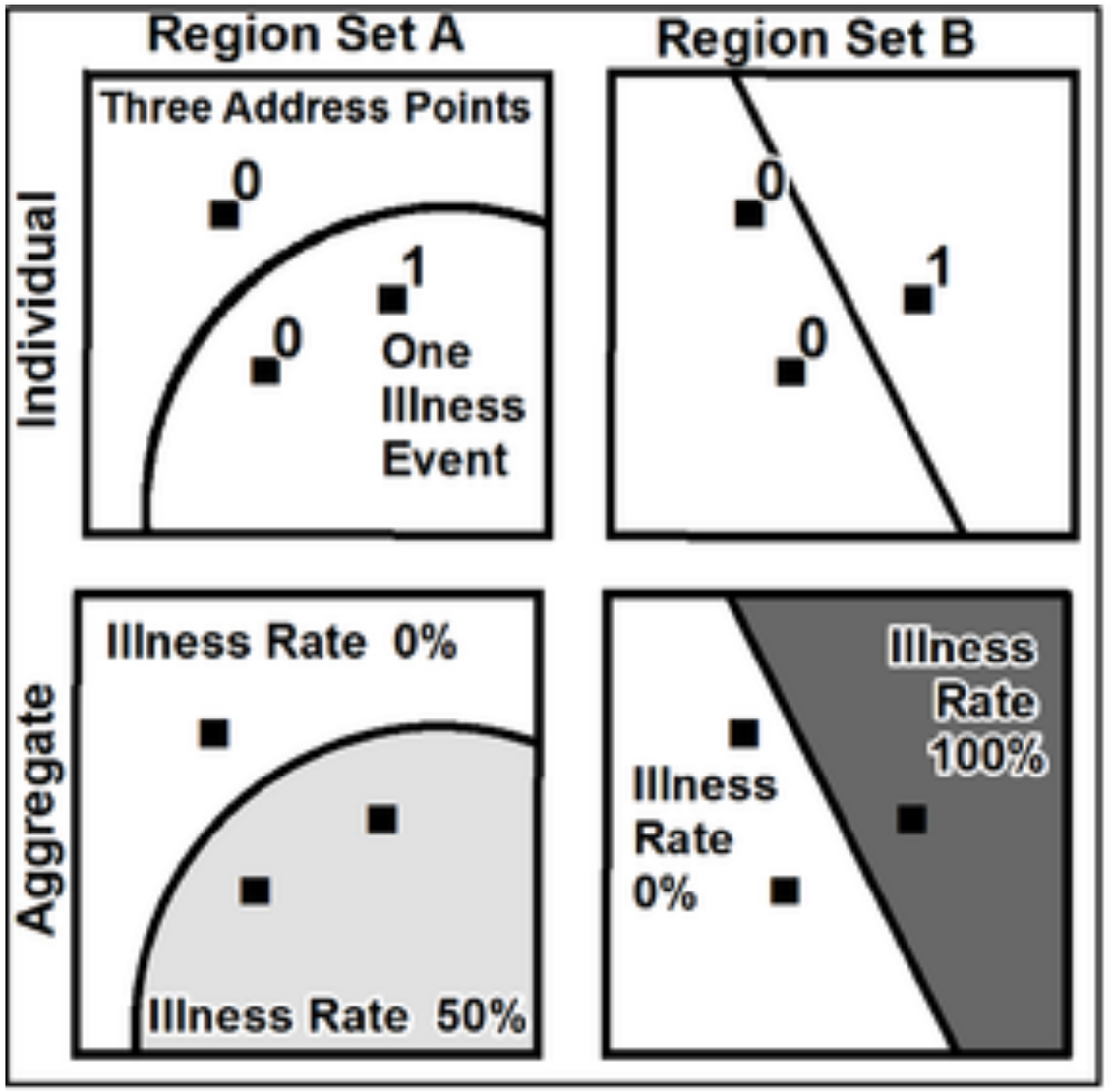
Individuals, shops,..



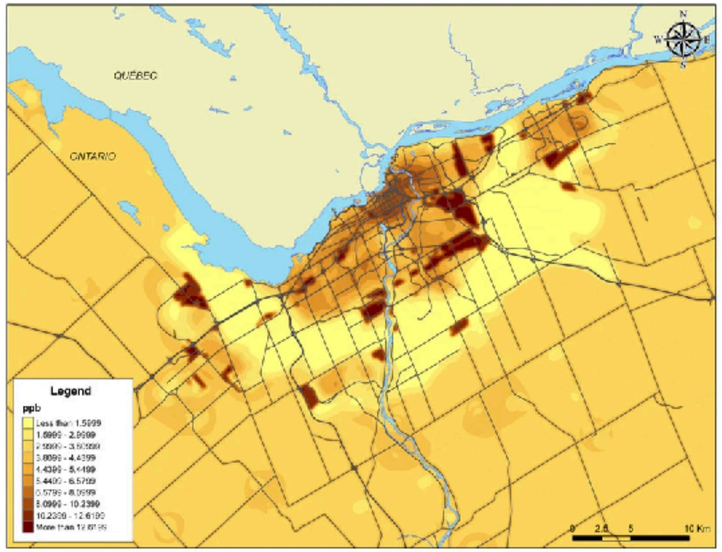
Districts, regions,…

Openshaw, 1984

A common source of bias in spatial aggregation: MAUP

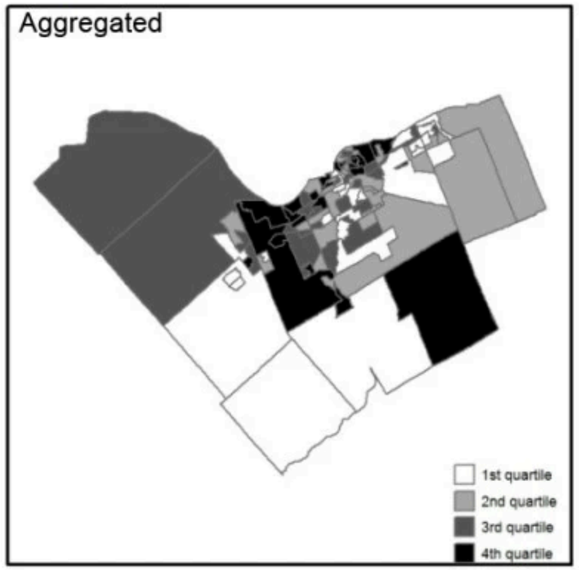


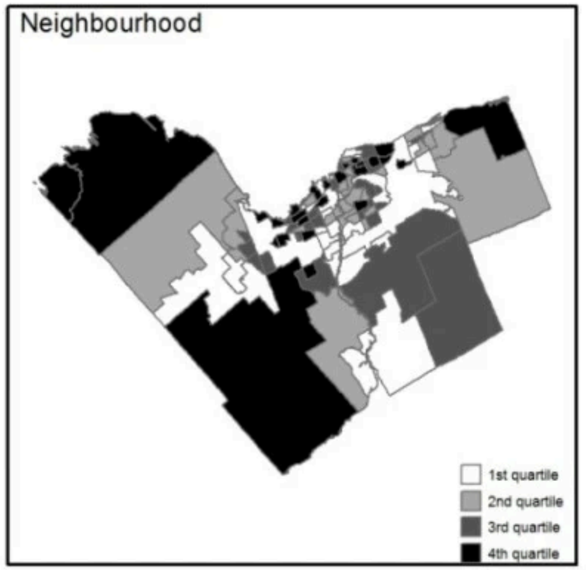
A common source of bias in spatial aggregation: MAUP

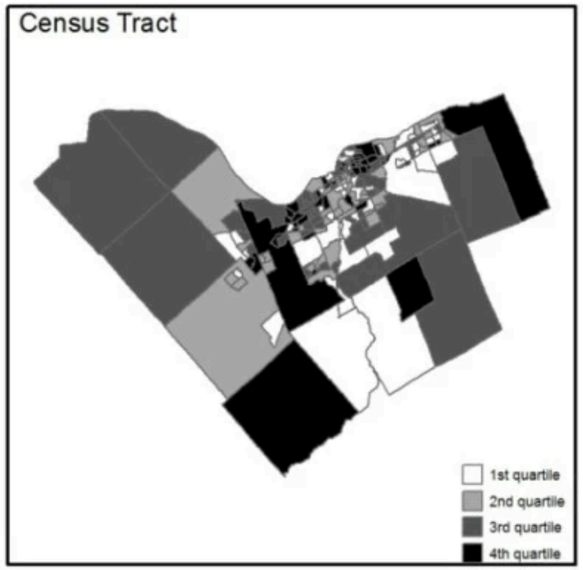


Pollution

Morbidity

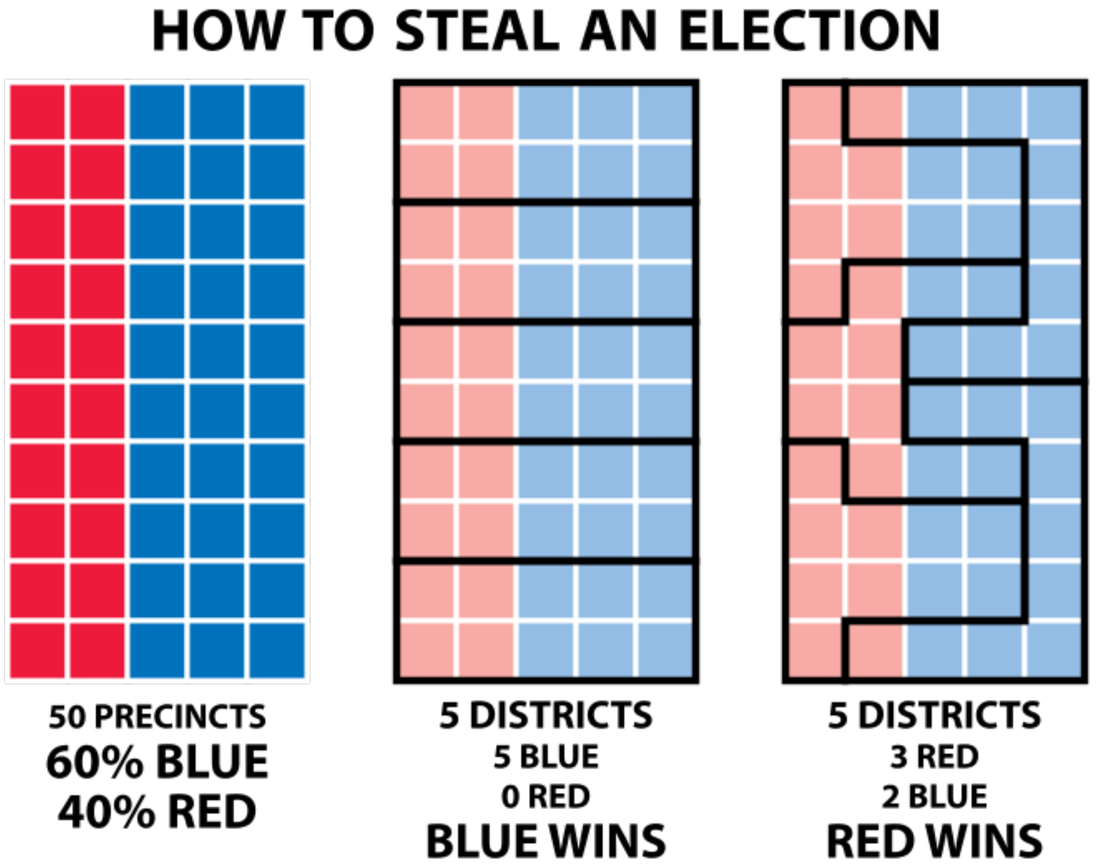


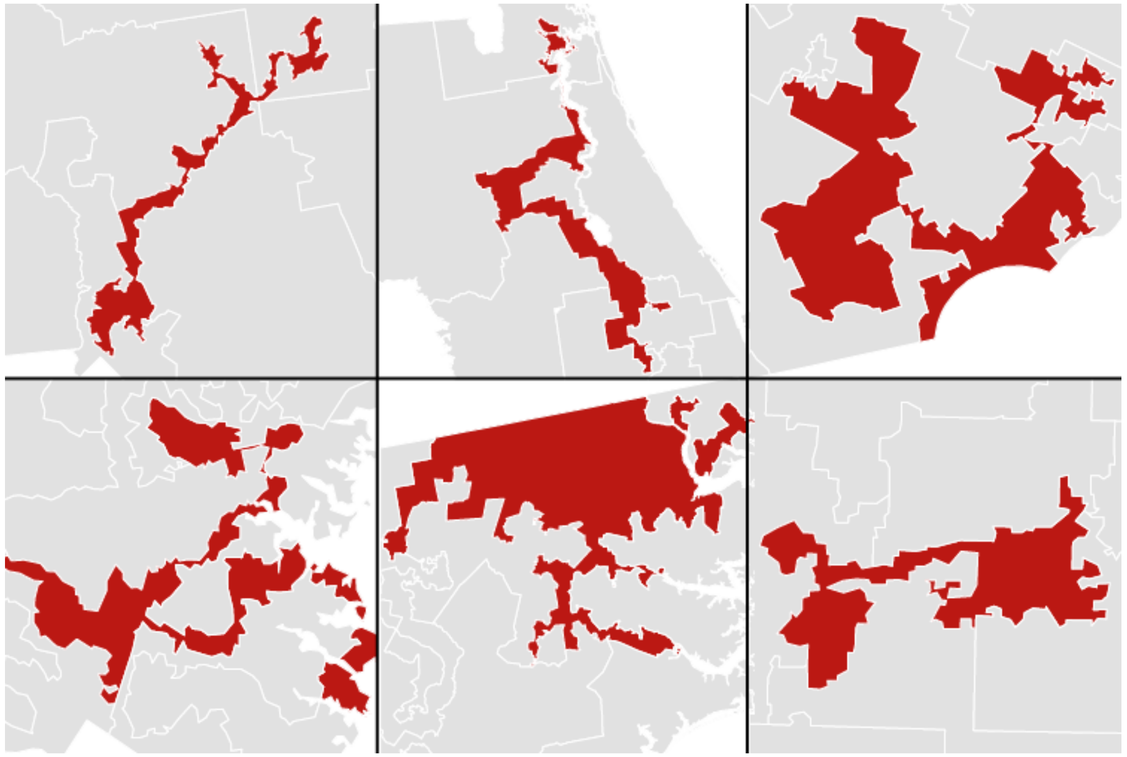




https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-10-58

The MAUP is abused for Gerrymandering

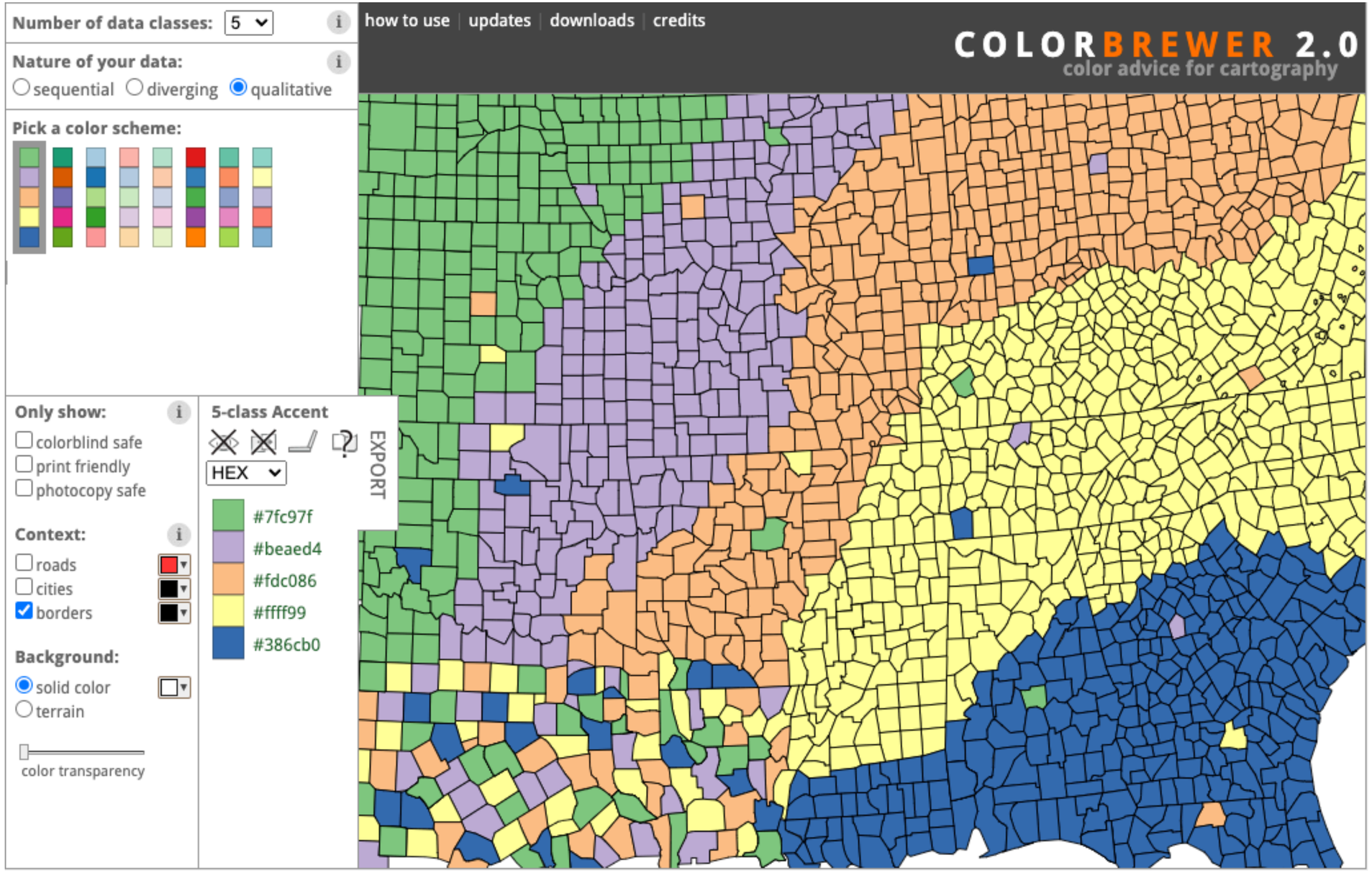




https://youtu.be/E2EnuHsRJd4?t=504

Always mind the MAUP when exploring aggregated data

2) Colors


Number of data classes: 5
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
5-class Accent
HEX
EXPORT
#7fc97f
#beaed4
#fdc086
#ffff99
#386cb0
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography

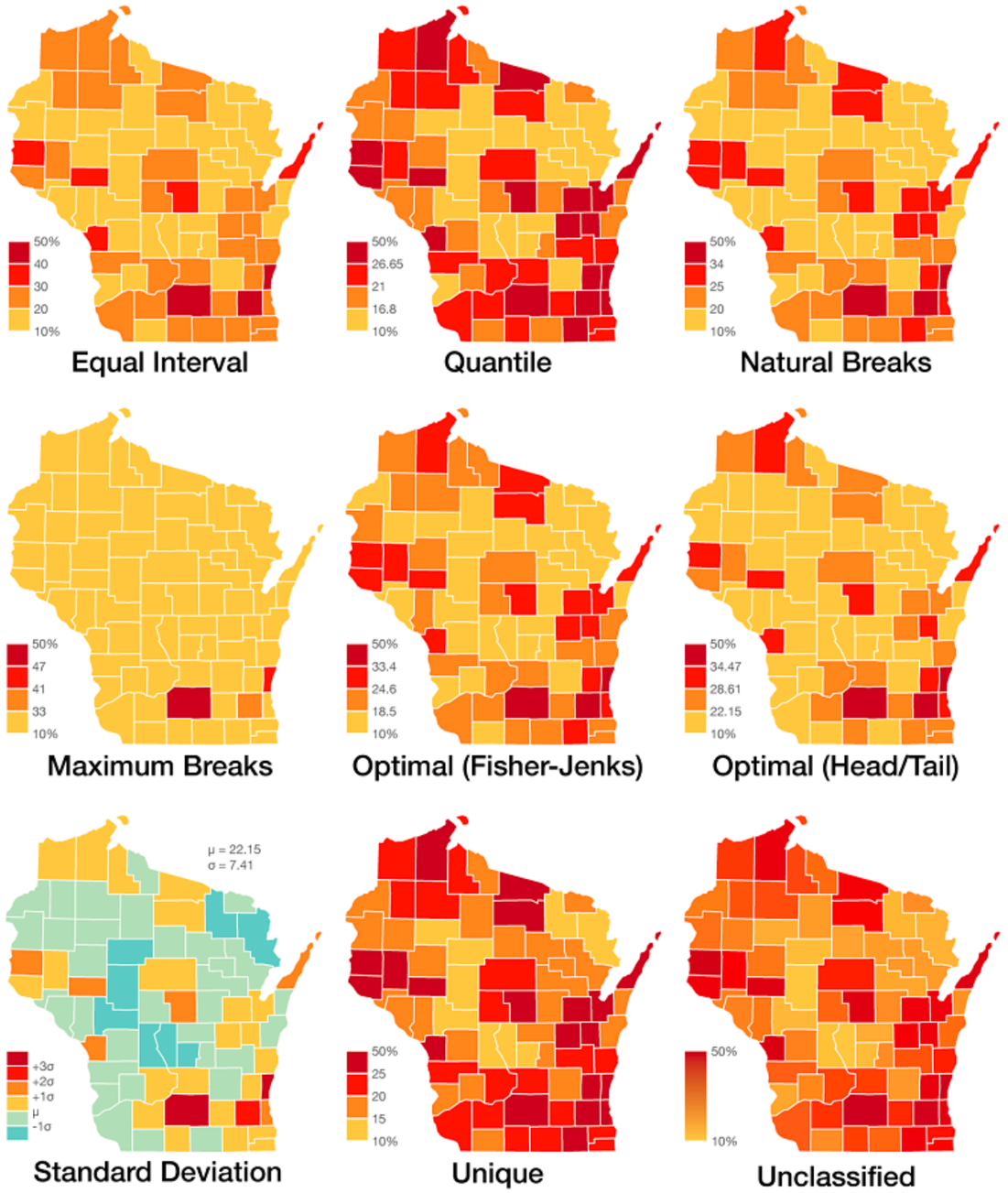
3) Classes

Classification: Grouping similar phenomena to gain relative simplicity in communication and interpretation. Also known as **binning**.

Data can be classified very differently

Same data,
Same spatial units,
Same colors,
Different classification

Percentage of residents over 25 with a Bachelor's degree



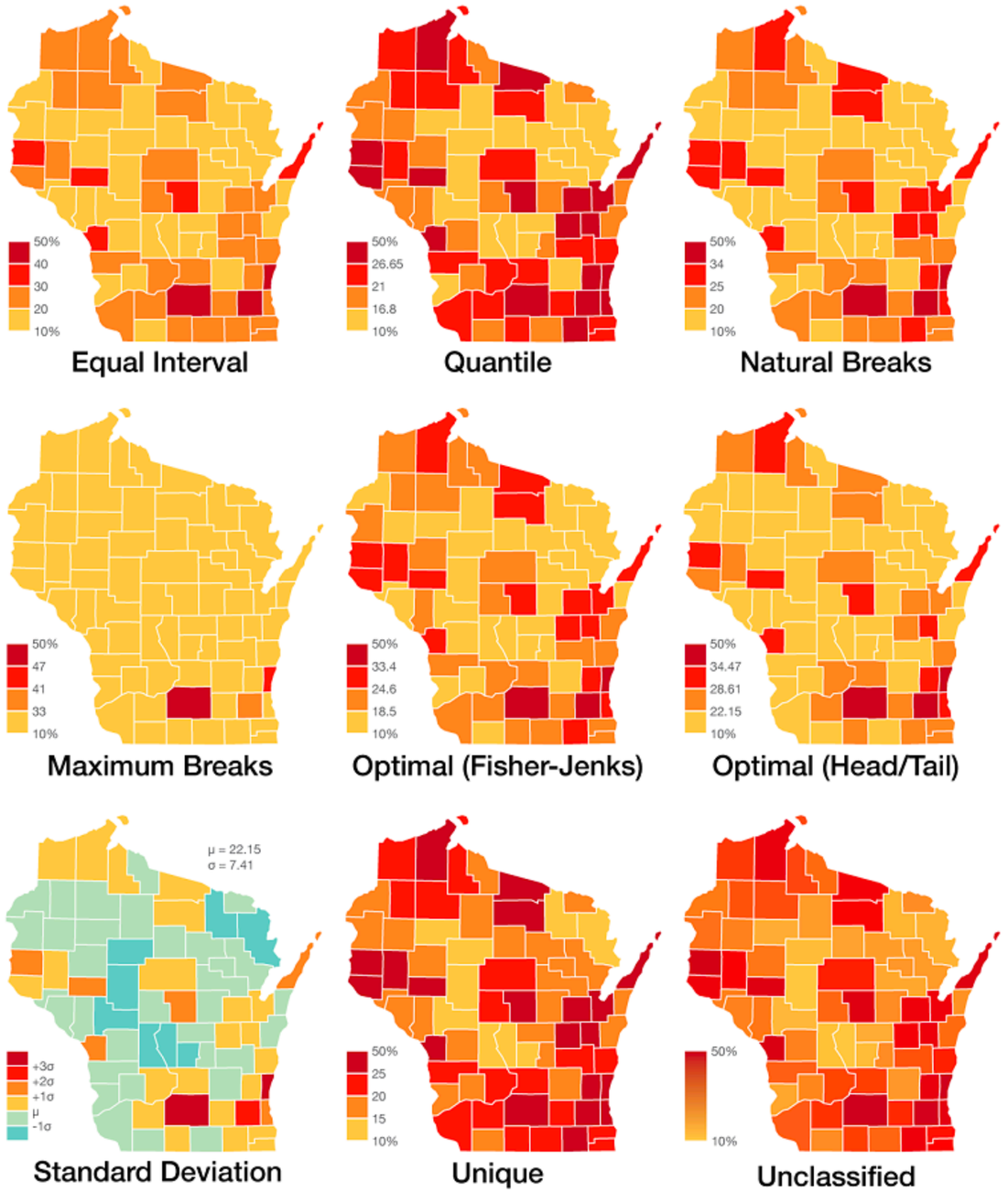
https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Data can be classified very differently

A data set does not have a “perfect" choropleth map

As with MAUP, you can create *arbitrary* interpretations of the same data!

Percentage of residents over 25 with a Bachelor's degree



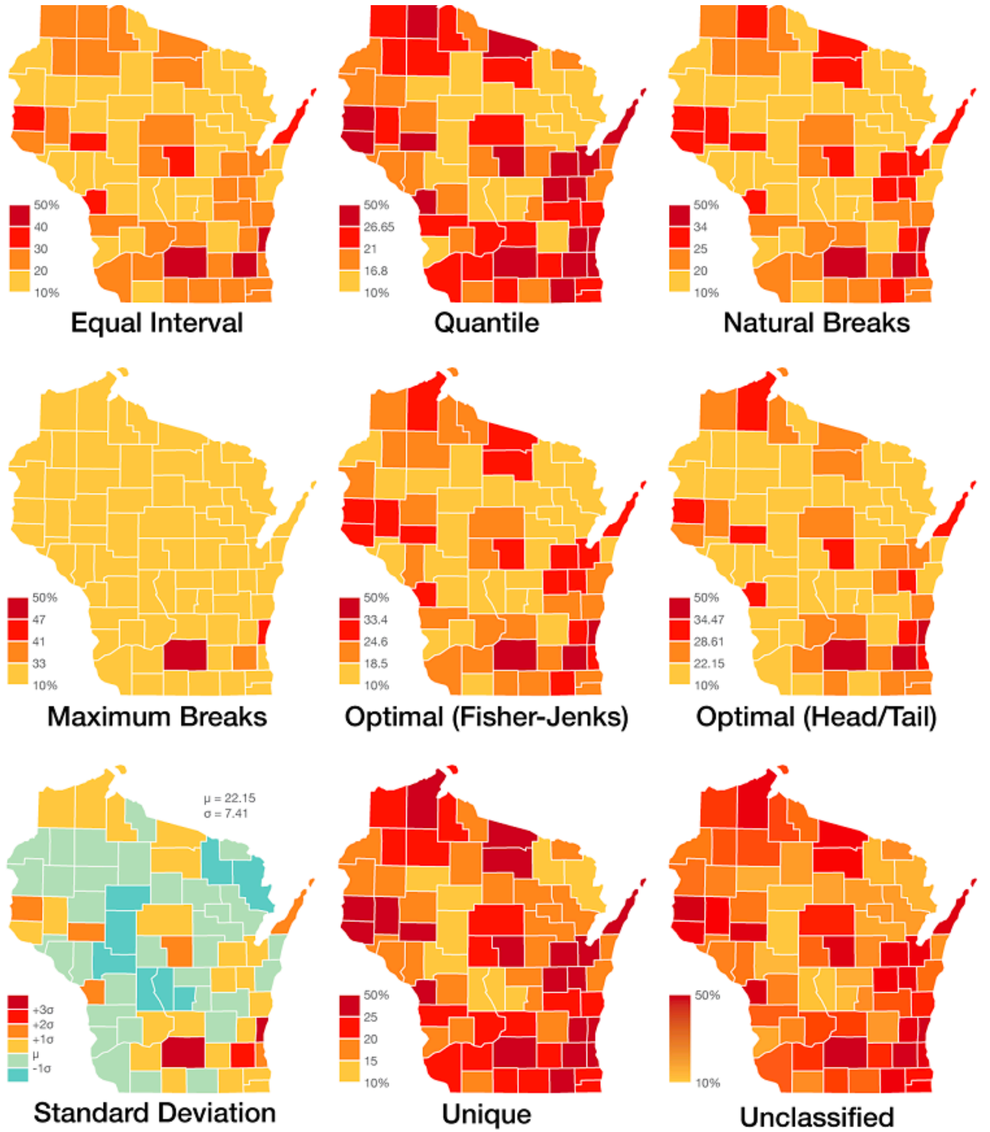
https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

The choice of class boundaries defines the classification scheme

mapclassify:

```
  'EqualInterval',
  'FisherJenks',
…
  'HeadTailBreaks',
  'JenksCaspall',
…
  'MaximumBreaks',
  'NaturalBreaks',
  'Percentiles',
…
  'Quantiles',
  'StdMean',
  'UserDefined',
```

Percentage of residents over 25 with a Bachelor's degree

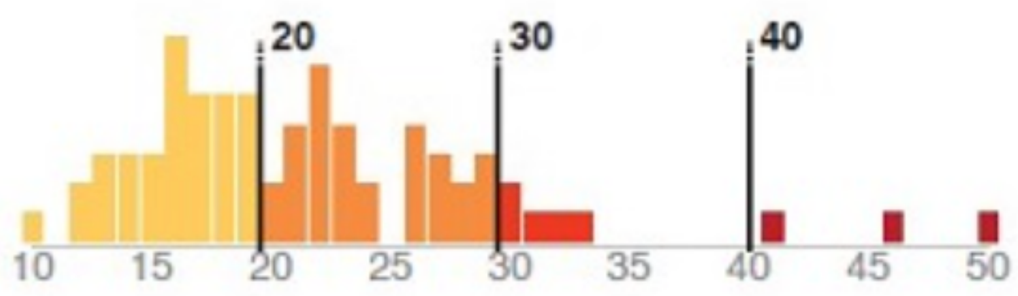


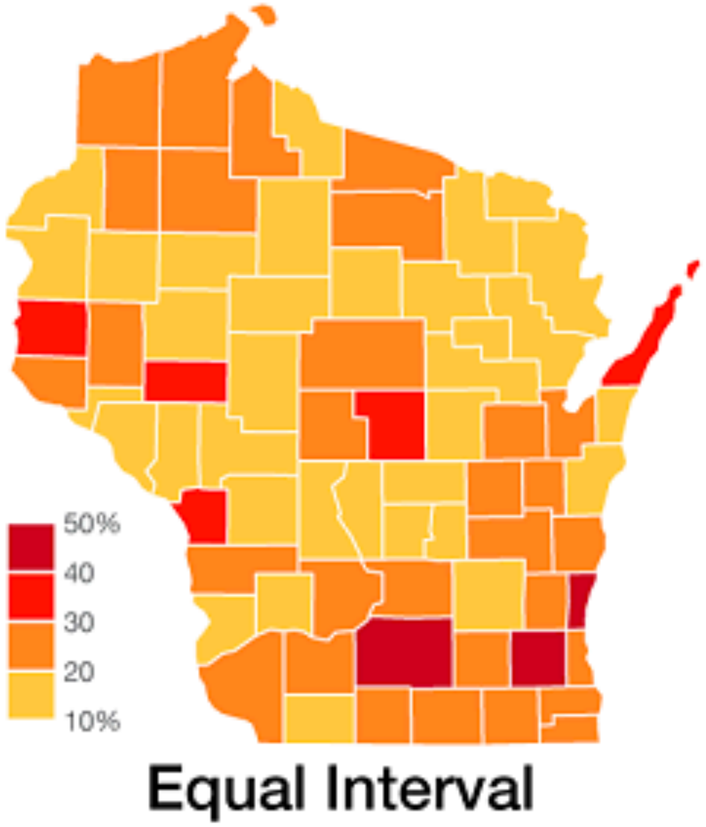
https://pypi.org/project/mapclassify/

https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Equal Interval

For uniformly distributed data with familiar data ranges.

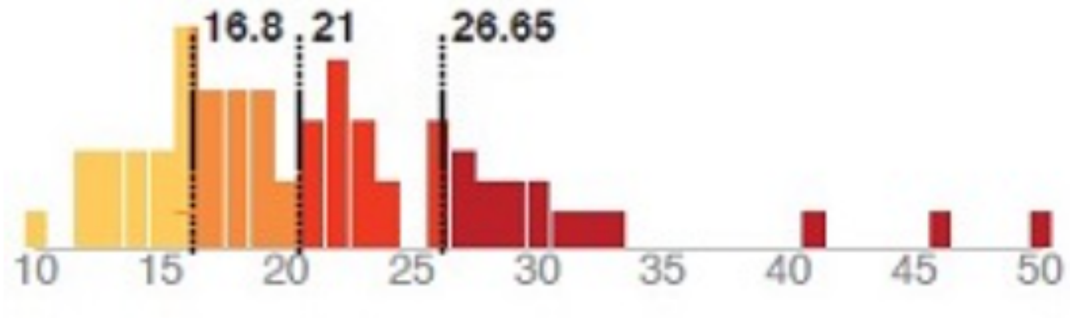


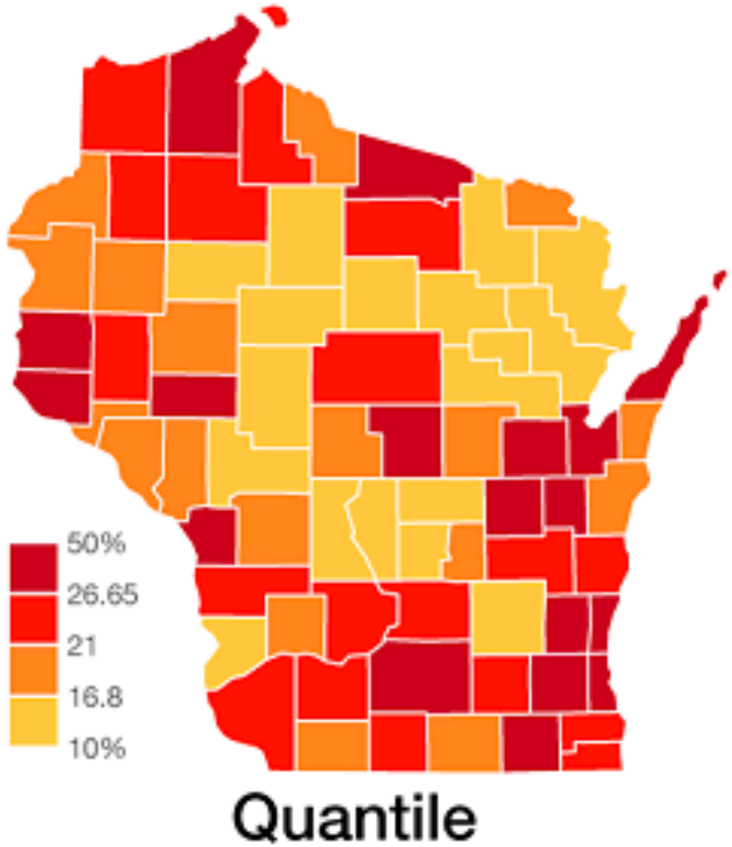


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Quantiles (Equal Count)

For evenly distributed data and ordinal data

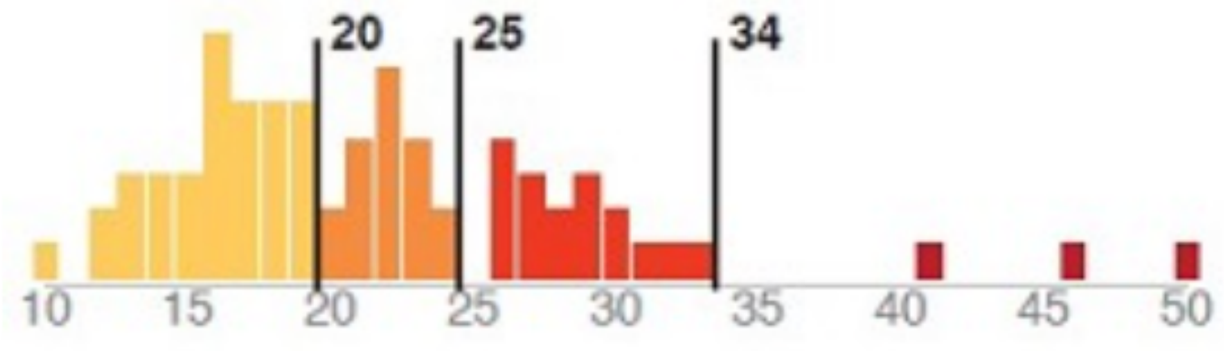


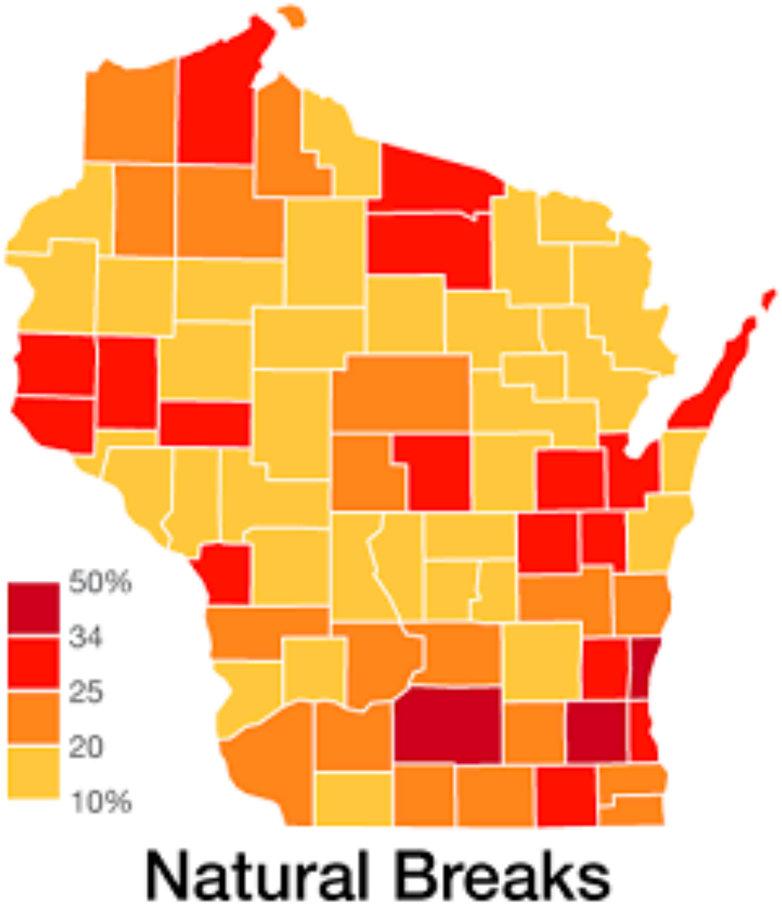


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Natural Breaks

For clustered data

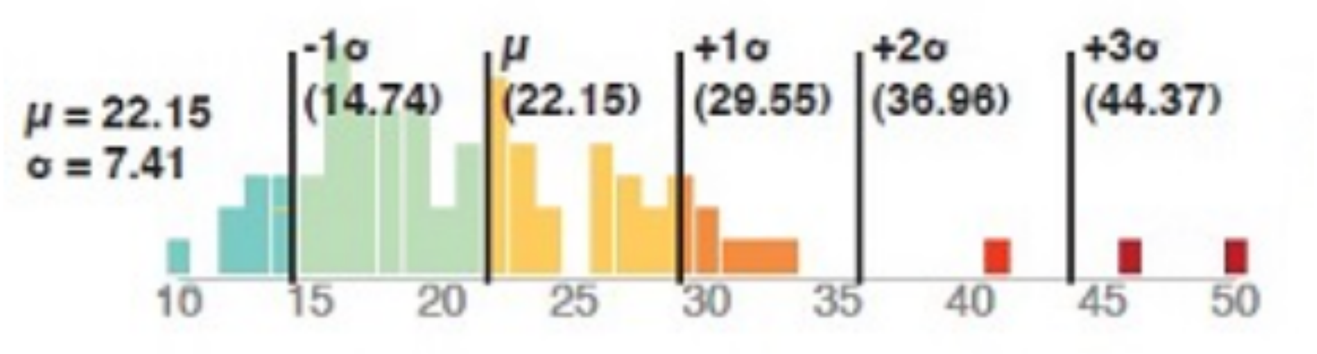


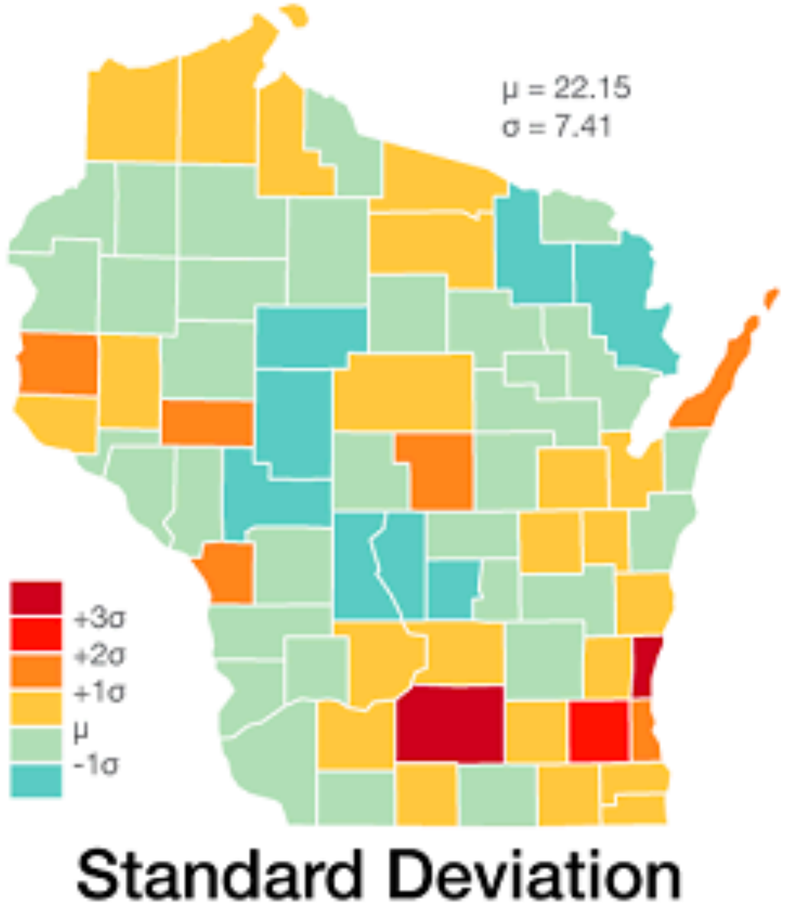


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Mean-Standard Deviation

For normally distributed data

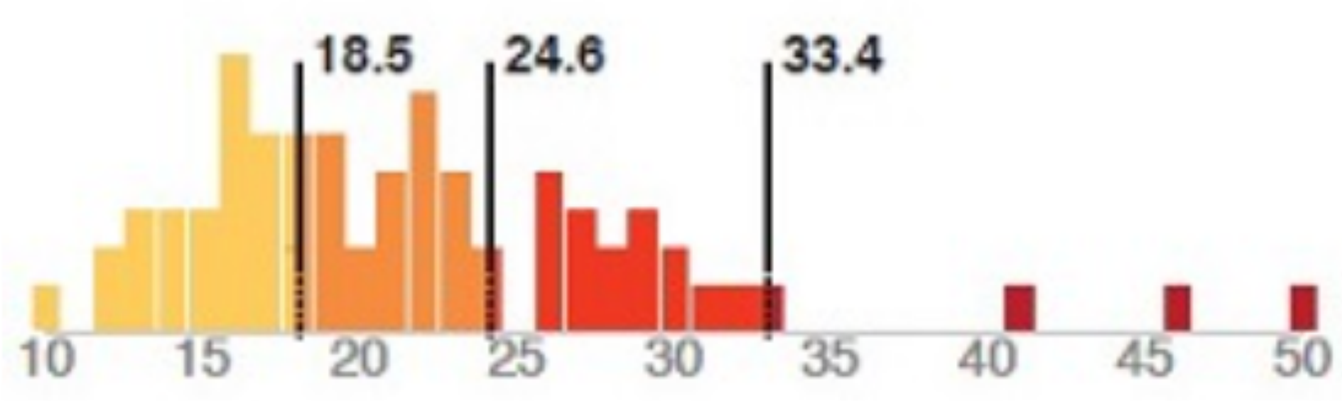


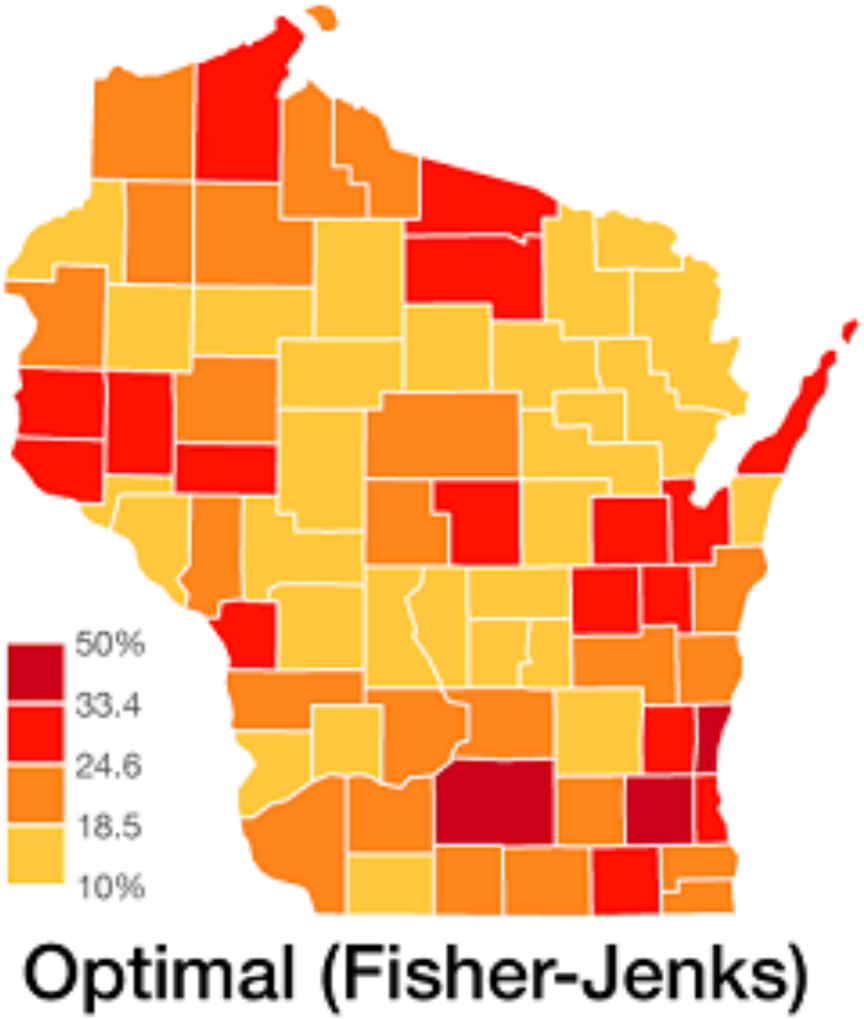


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Jenks-Caspall & Fisher-Jenks

For clustered and skewed data

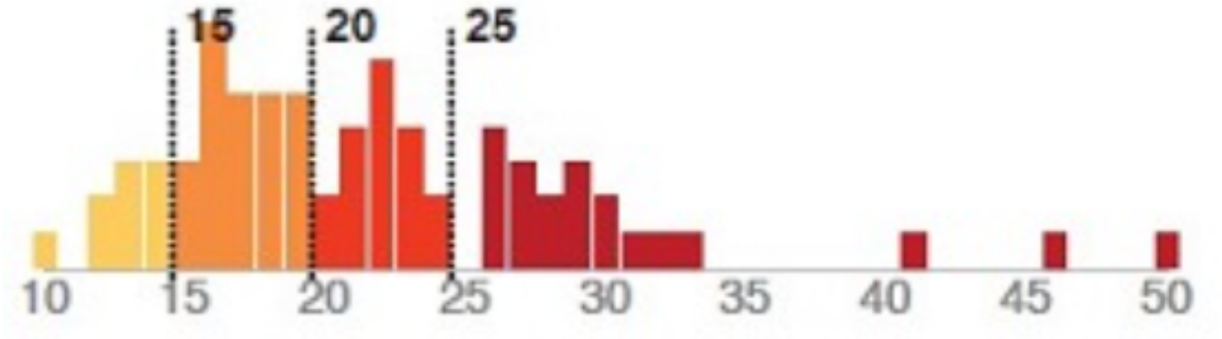


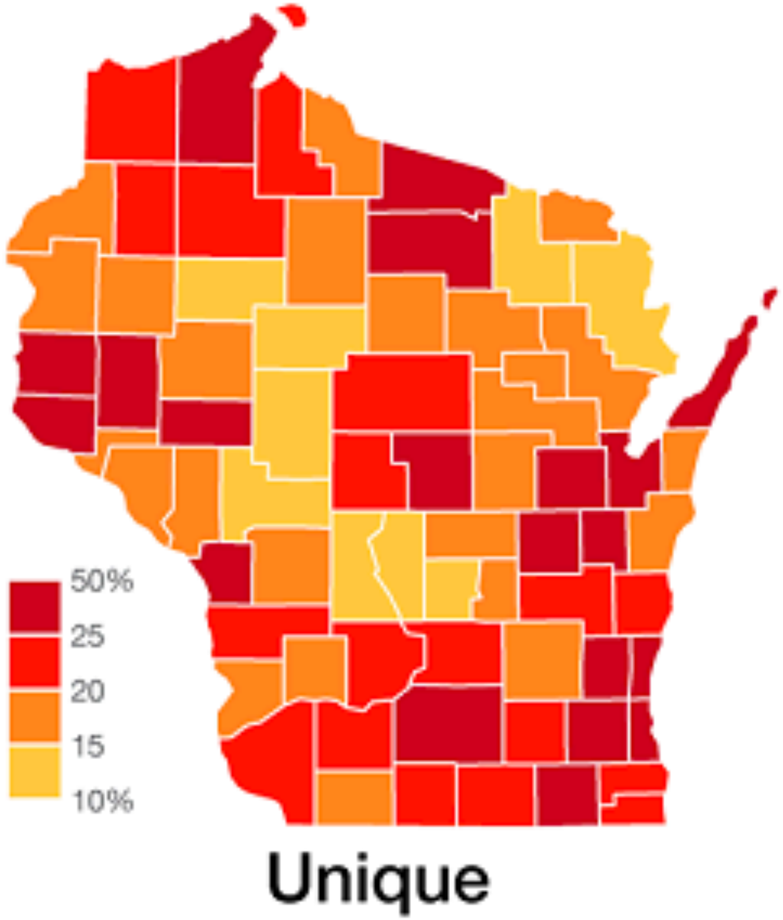


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Unique/Manual

For data where key numbers are important





https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Further materials

part2/part2spatialstatistics.ipynb
(at the bottom)

Spatial weights formalize: How do things relate in space?

> Everything is related to everything else, but near things are more related than distant things.
>
> Tobler's 1st law of geography

PySal: Python Spatial Analysis Library

Core library for geospatial analysis

https://pysal.org/

How are objects related to each other in space?

The spatial weight matrix W encodes the spatial relation between N objects

$$W = \begin{pmatrix} 0 & w_{12} & \dots & w_{1N} \\ w_{21} & \ddots & w_{ij} & \vdots \\ \vdots & w_{ji} & 0 & \vdots \\ w_{N1} & \dots & \dots & 0 \end{pmatrix}$$

N times N, positive

$w_{ii} = 0$

Generally, all non-zero elements in a row *i* are called the neighbors of object *i*. How to define ‘neighbor’?

How are objects related to each other in space?

The spatial weight matrix W encodes the spatial relation between N objects

$$W = \begin{pmatrix} 0 & w_{12} & \dots & w_{1N} \\ w_{21} & \ddots & w_{ij} & \vdots \\ \vdots & w_{ji} & 0 & \vdots \\ w_{N1} & \dots & \dots & 0 \end{pmatrix}$$

N times N, positive

$w_{ii} = 0$

Contiguity Is object 2 "next to" object 1?

How are objects related to each other in space?

The spatial weight matrix W encodes the spatial relation between N objects

$$W = \begin{pmatrix} 0 & w_{12} & \dots & w_{1N} \\ w_{21} & \ddots & w_{ij} & \vdots \\ \vdots & w_{ji} & 0 & \vdots \\ w_{N1} & \dots & \dots & 0 \end{pmatrix}$$

N times N, positive

$$w_{ii} = 0$$

Contiguity — Is object 2 "next to" object 1?

Distance — Is object 2 "close" to object 1?

How are objects related to each other in space?

The spatial weight matrix W encodes the spatial relation between N objects

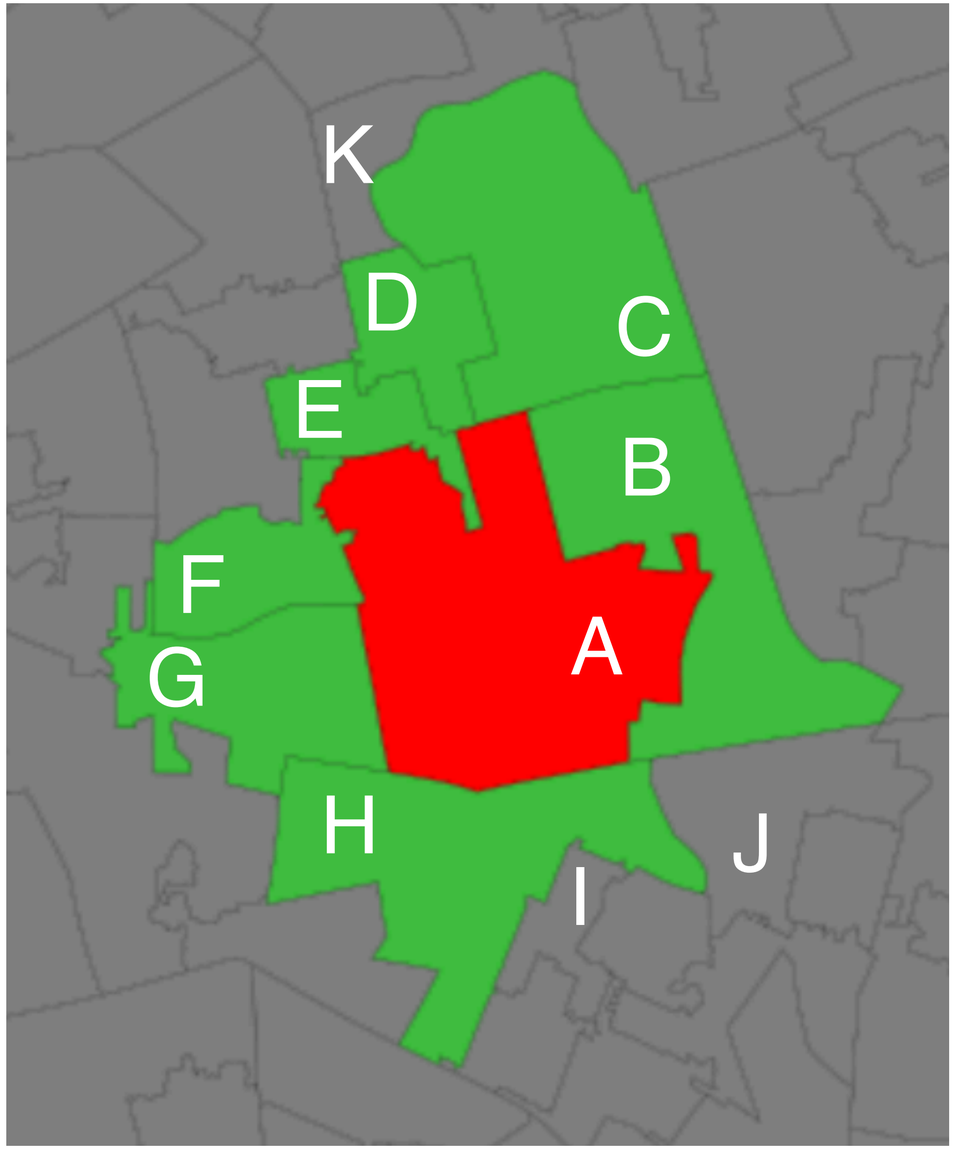
$$W = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1N} \\ w_{21} & \ddots & w_{ij} & \vdots \\ \vdots & w_{ji} & 0 & \vdots \\ w_{N1} & \cdots & \cdots & 0 \end{pmatrix}$$

N times N, positive

$w_{ii} = 0$

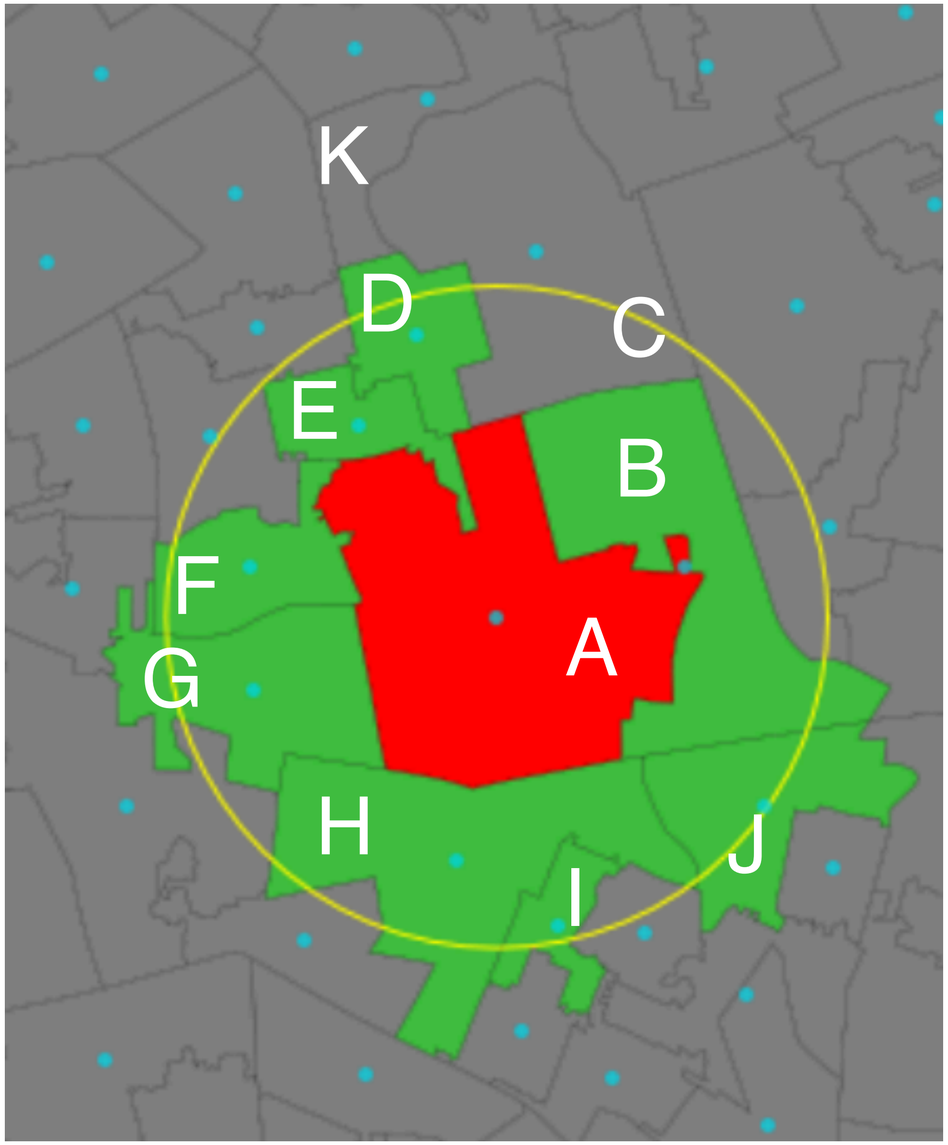
Contiguity	Is object 2 "next to" object 1?
Distance	Is object 2 "close" to object 1?
Block	Is object 2 in the same "place" as object 1?

Contiguity-based W: sharing a boundary



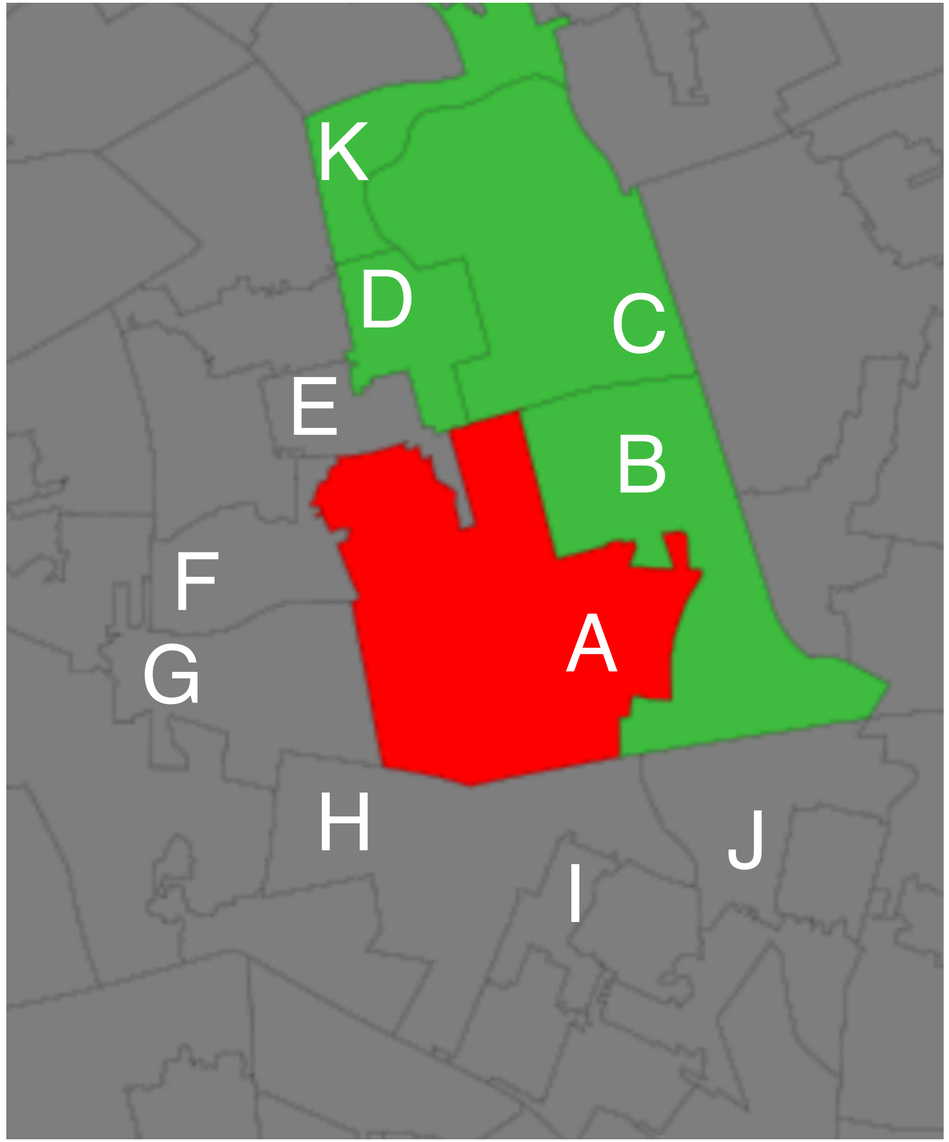
	A	B	C	D	E	F	G	H	I	J	K
A	0	1	1	1	1	1	1	1	0	0	0
B	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...

Distance-based W: being closer than a threshold



	A	B	C	D	E	F	G	H	I	J	K
A	0	1	0	1	1	1	1	1	1	1	0
B	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...

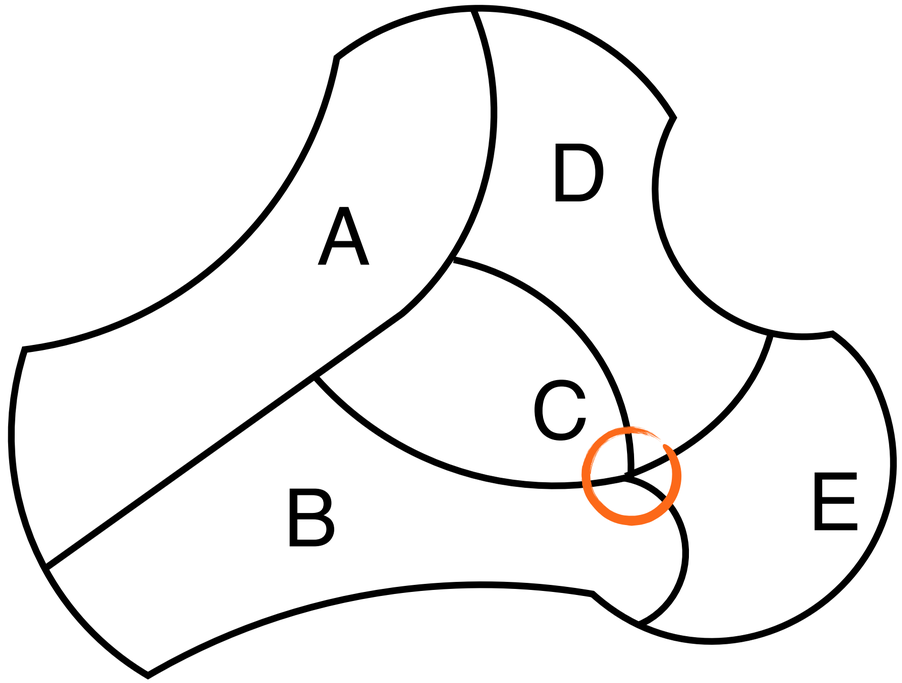
Block-based W: being in the same administrative unit



	A	B	C	D	E	F	G	H	I	J	K
A	0	1	1	1	0	0	0	0	0	0	1
B	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...

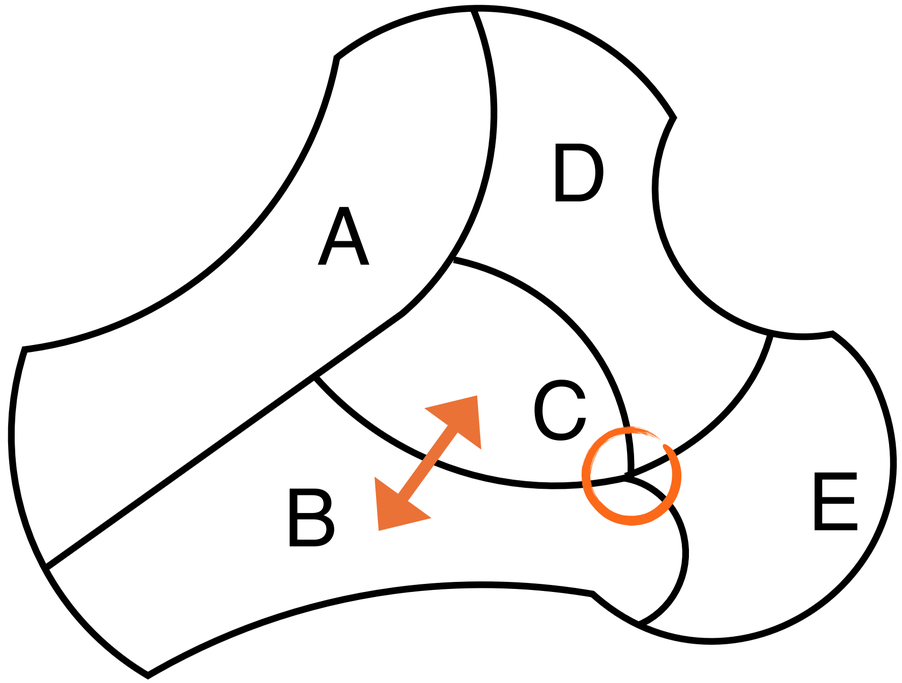
Special cases and variations

Contiguity-based W: What is a boundary?



	A	B	C	D	E
A	0	1	1	1	0
B	1	0	1	?	1
C	1	1	0	1	?
D	1	?	1	0	1
E	0	1	?	1	0

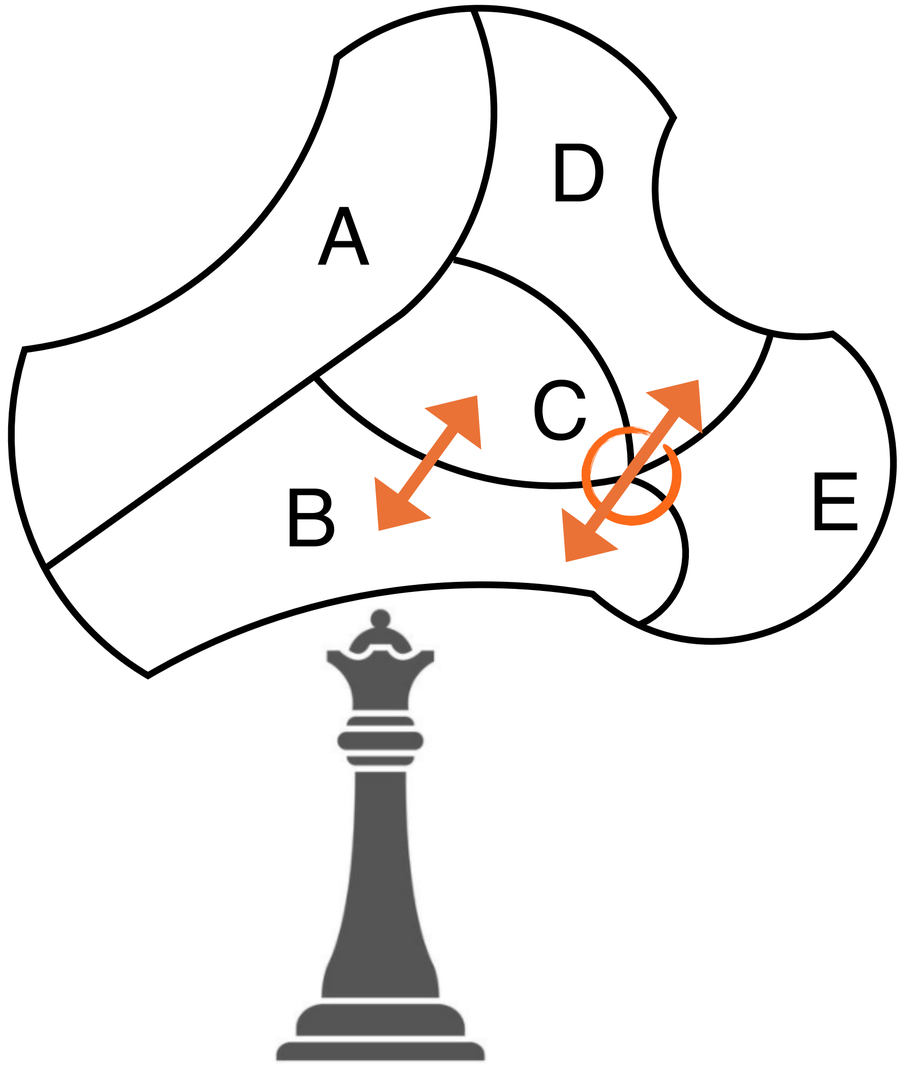
Contiguity-based W: What is a boundary?



ROOK

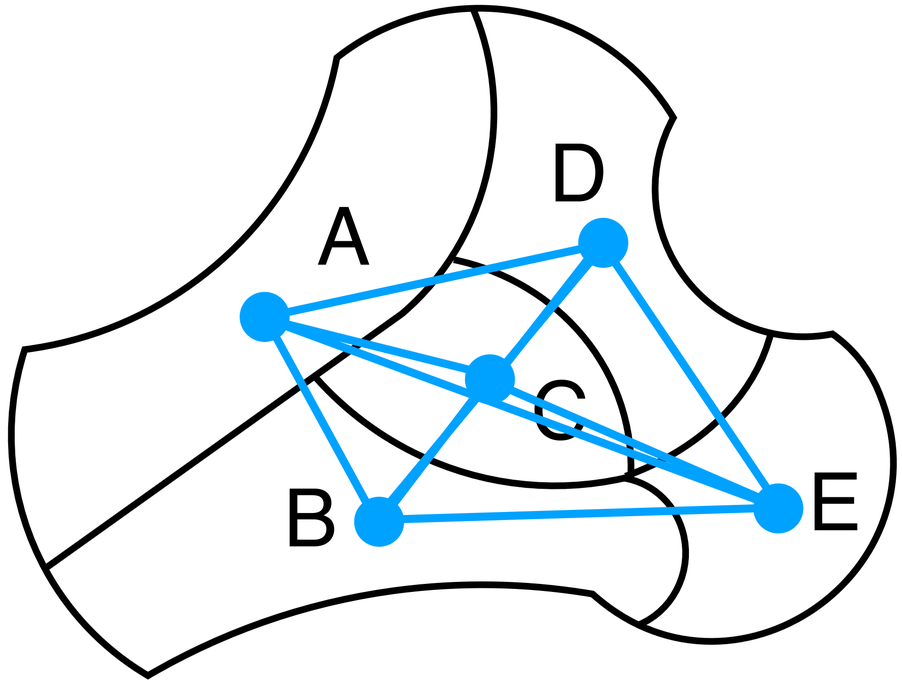
	A	B	C	D	E
A	0	1	1	1	0
B	1	0	1	0	1
C	1	1	0	1	0
D	1	0	1	0	1
E	0	1	0	1	0

Contiguity-based W: What is a boundary?



	A	B	C	D	E
A	0	1	1	1	0
B	1	0	1	1	1
C	1	1	0	1	1
D	1	1	1	0	1
E	0	1	1	1	0

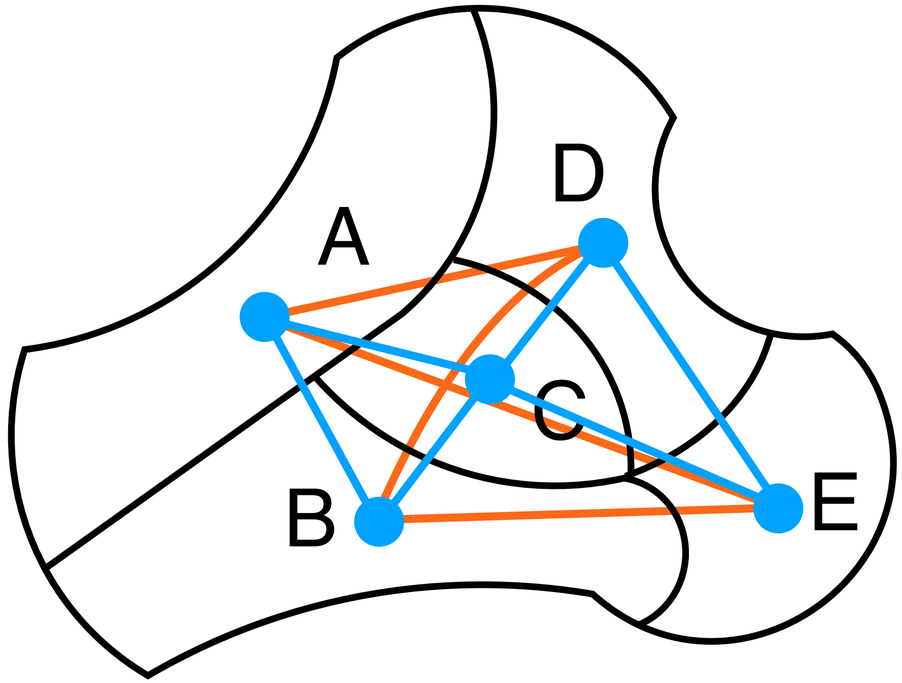
Distance-based W: Values can be continuous



$$w_{ij} = \frac{1}{d_{ij}^{\alpha}}$$

	A	B	C	D	E
A	0	0.58	0.57	0.39	0.25
B	0.58	0	0.76	0.38	0.33
C	0.57	0.76	0	0.76	0.43
D	0.39	0.38	0.76	0	0.42
E	0.25	0.33	0.43	0.42	0

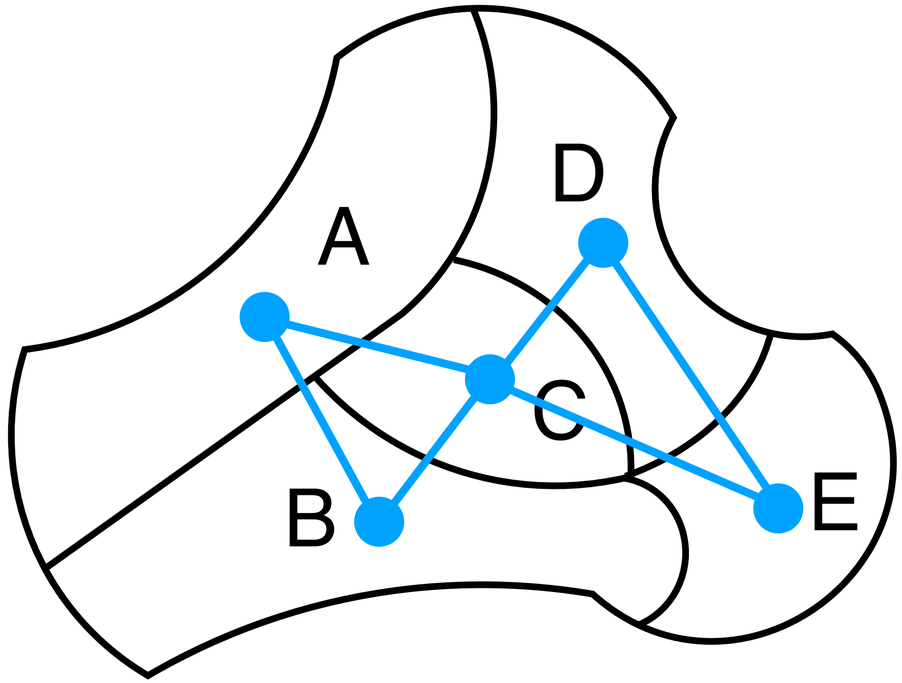
Distance-based W: Values can be continuous



We want a threshold
to keep W sparse!

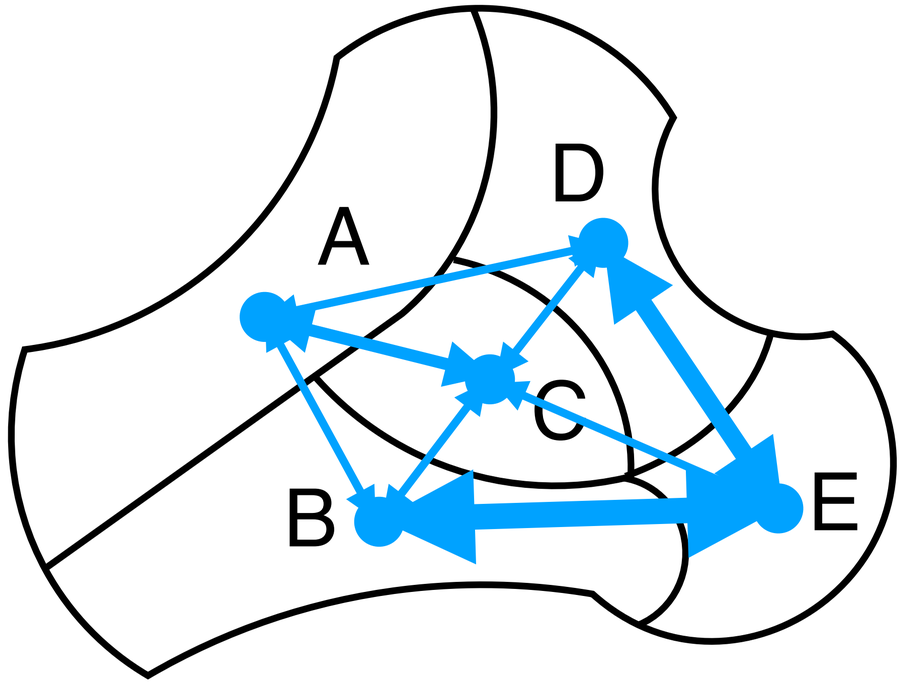
t=0.4	A	B	C	D	E
A	0	0.58	0.57	0	0
B	0.58	0	0.76	0	0
C	0.57	0.76	0	0.76	0.43
D	0	0	0.76	0	0.42
E	0	0	0.43	0.42	0

Distance-based W: KNN (k nearest neighbors)



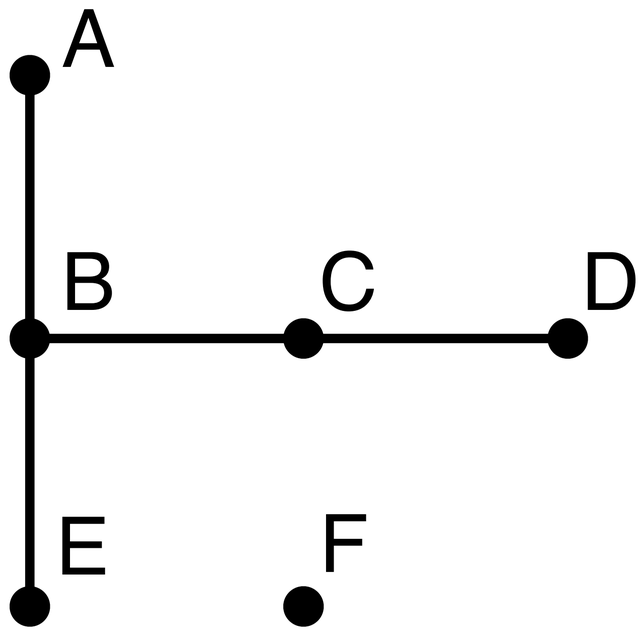
k=2	A	B	C	D	E
A	0	1	1	0	0
B	1	0	1	0	0
C	0	1	0	1	0
D	0	0	1	0	1
E	0	0	1	1	0

Interaction-based W: Flows



	A	B	C	D	E
A	0	1	2	1	0
B	1	0	1	0	5
C	2	1	0	1	1
D	1	0	1	0	2
E	0	4	1	3	0

The structure can be a network

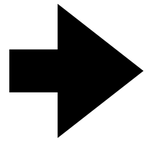


$$w_{ij} = \frac{1}{l_{ij}}$$

	A	B	C	D	E	F
A	0	1	0.5	0.33	0.5	0
B	1	0	1	0.5	1	0
C	0.5	1	0	1	0.5	0
D	0.33	0.5	1	0	0.33	0
E	0.5	1	0.5	0.33	0	0
F	0	0	0	0	0	0

It is common to standardize W: divide all by sum of row

	A	B	C	D	E
A	0	1	2	1	0
B	1	0	1	0	5
C	2	1	0	1	1
D	1	0	1	0	2
E	0	4	1	3	0



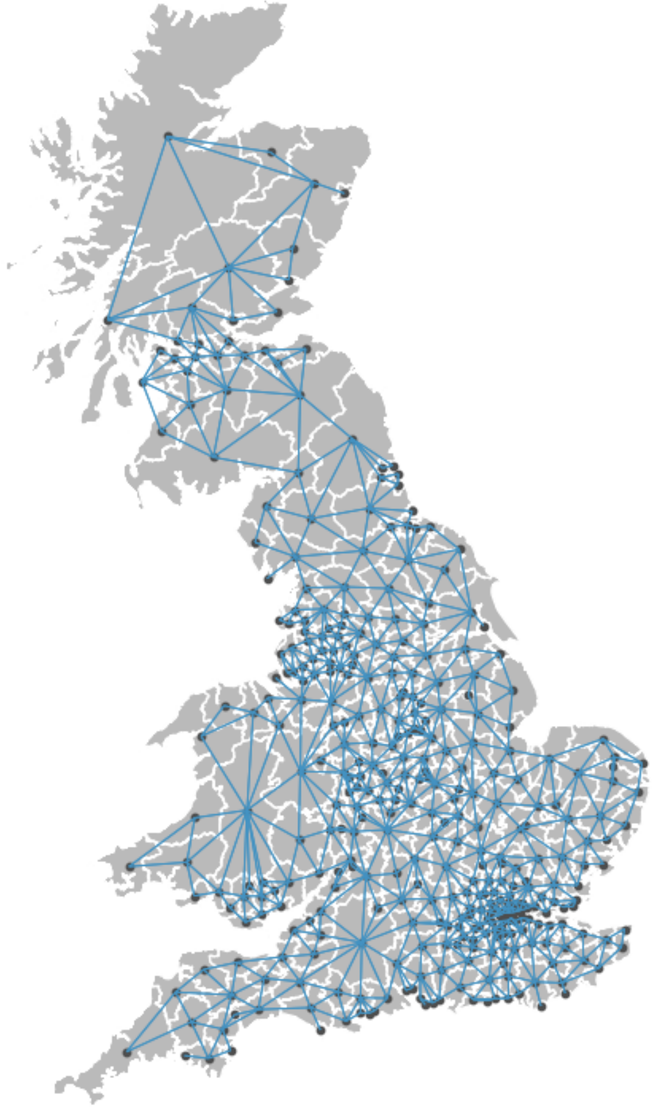
	A	B	C	D	E	Σ
A	0	1/4	1/2	1/4	0	1
B	1/7	0	1/7	0	5/7	1
C	2/5	1/5	0	1/5	1/5	1
D	1/4	0	1/4	0	1/2	1
E	0	1/2	1/8	3/8	0	1

The choice of W should reflect the studied interactions

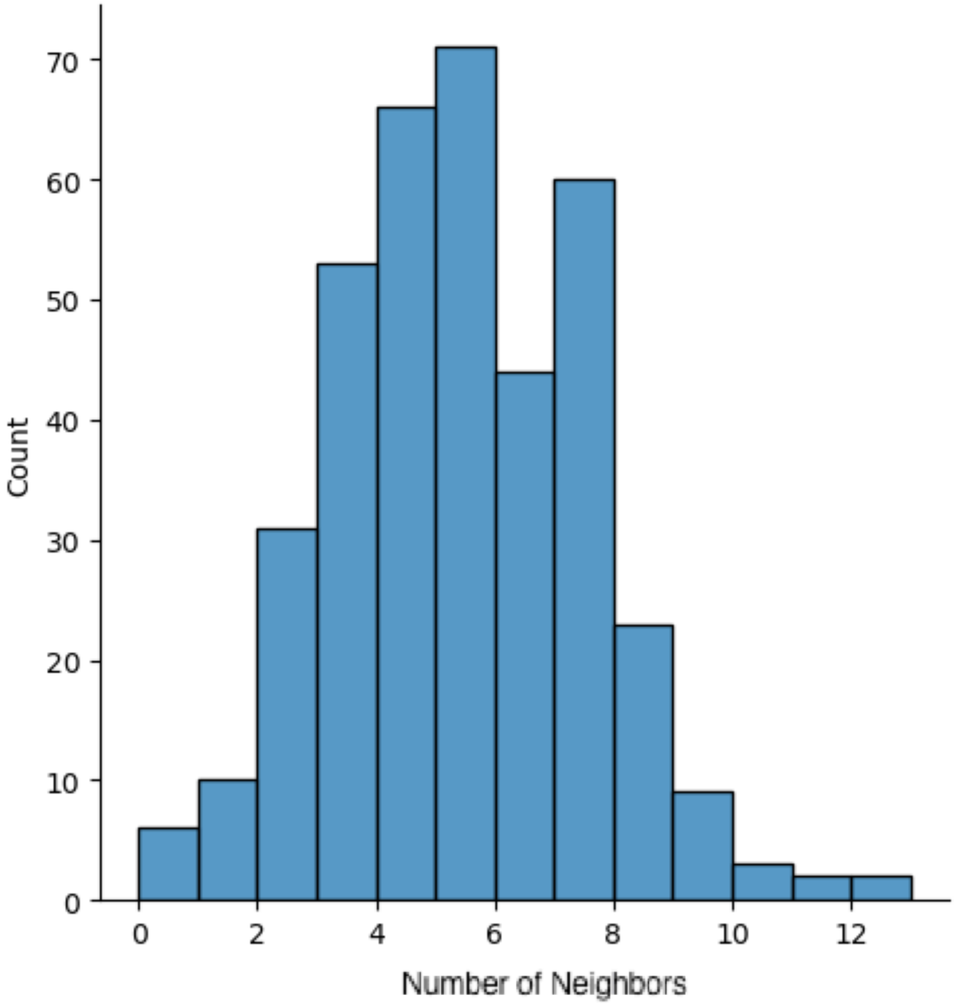
Spreading processes like a virus	→	Contiguity, Flow
Accessibility	→	Distance
Effects of laws on counties	→	Block

W defines a connectivity graph and histogram

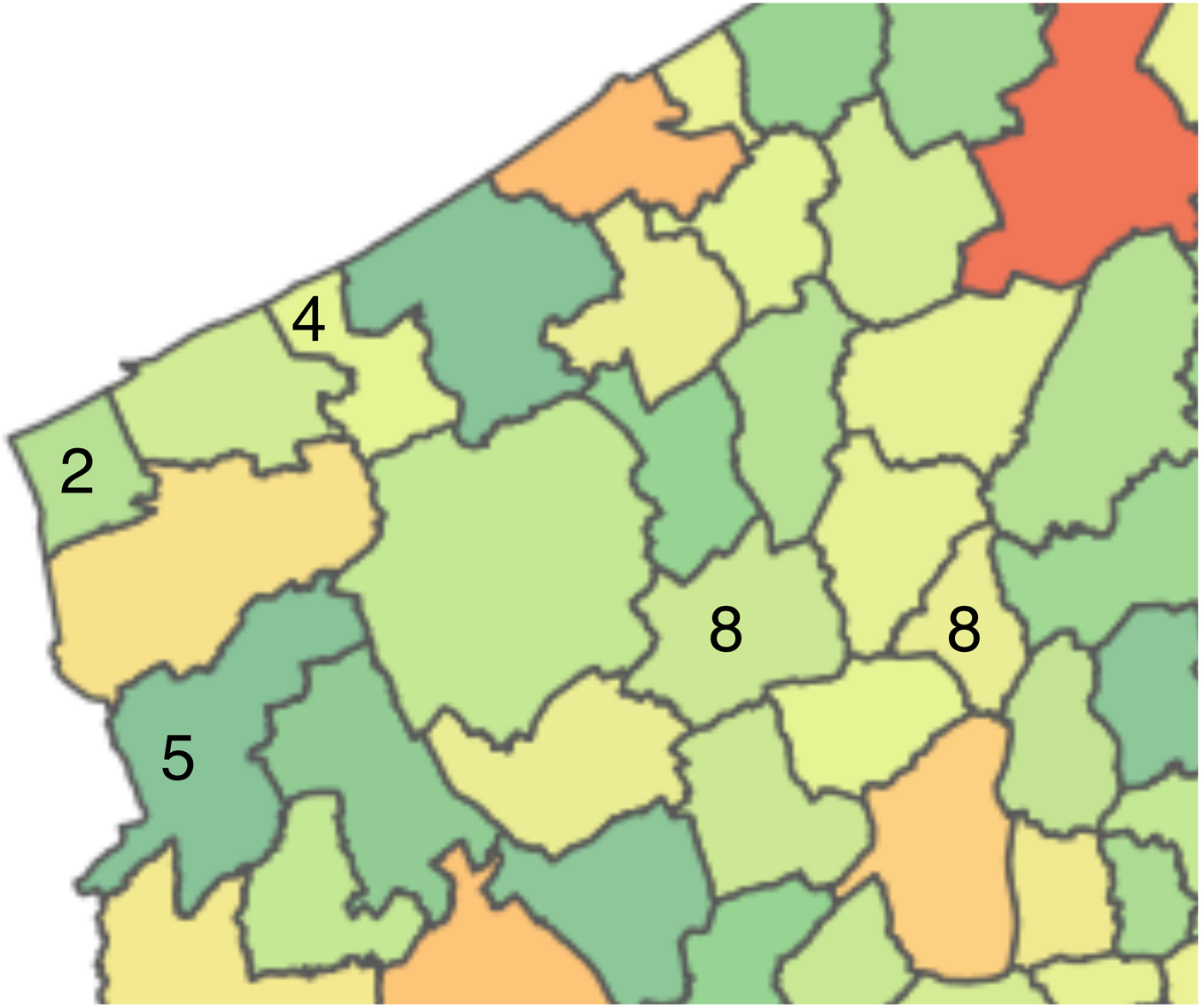
Connectivity graph



Connectivity histogram



It is important to consider edge effects!



Jupyter

part2/part2spatialstatistics.ipynb

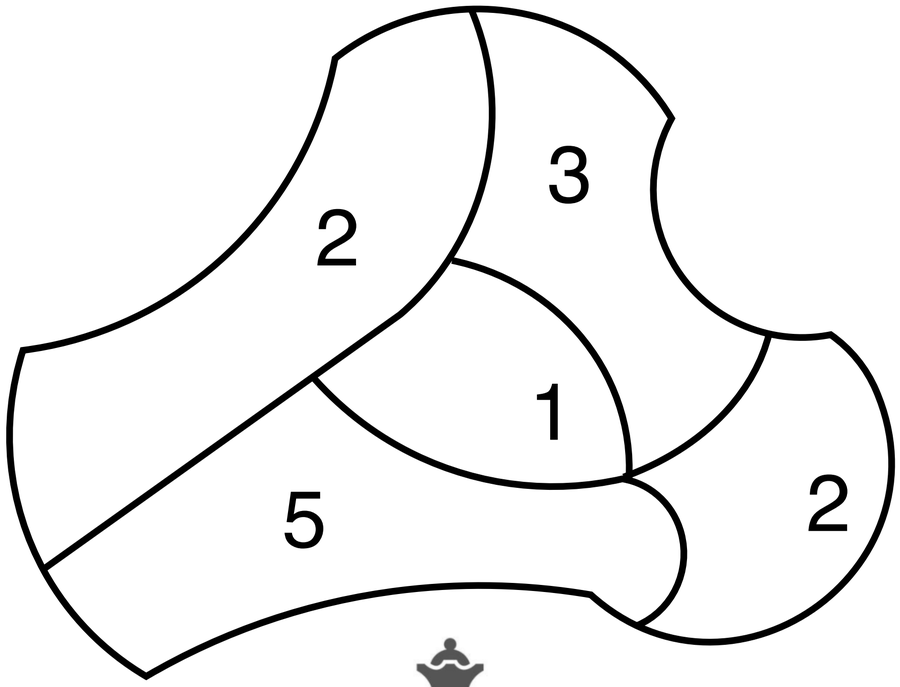
github.com/NERDSITU/gdstutorial

Spatial lag

The spatial lag is the weighted average value of neighbors

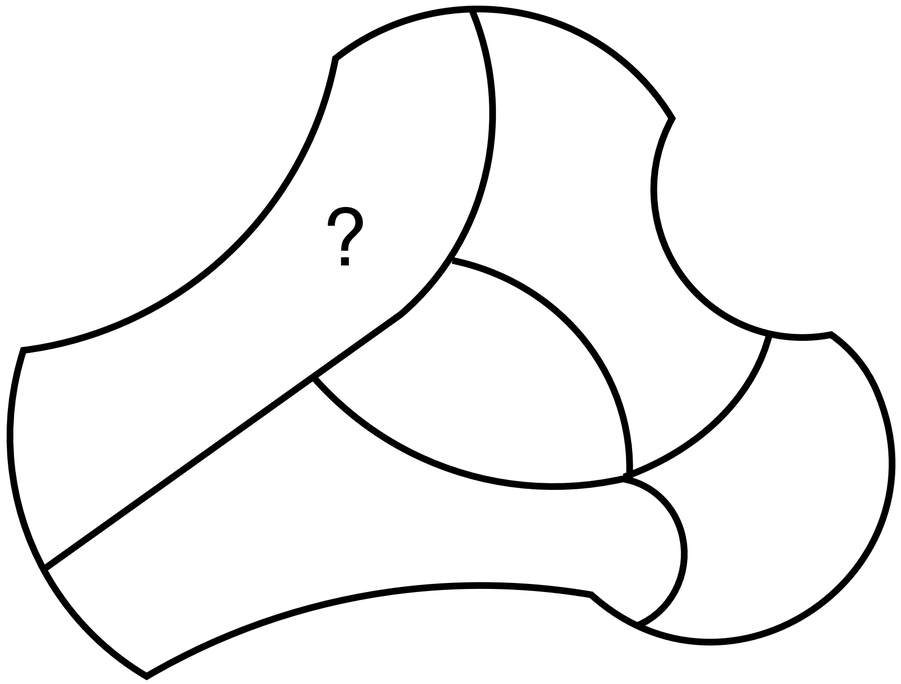
if W is standardized

Values $\mathbf{y}$



QUEEN

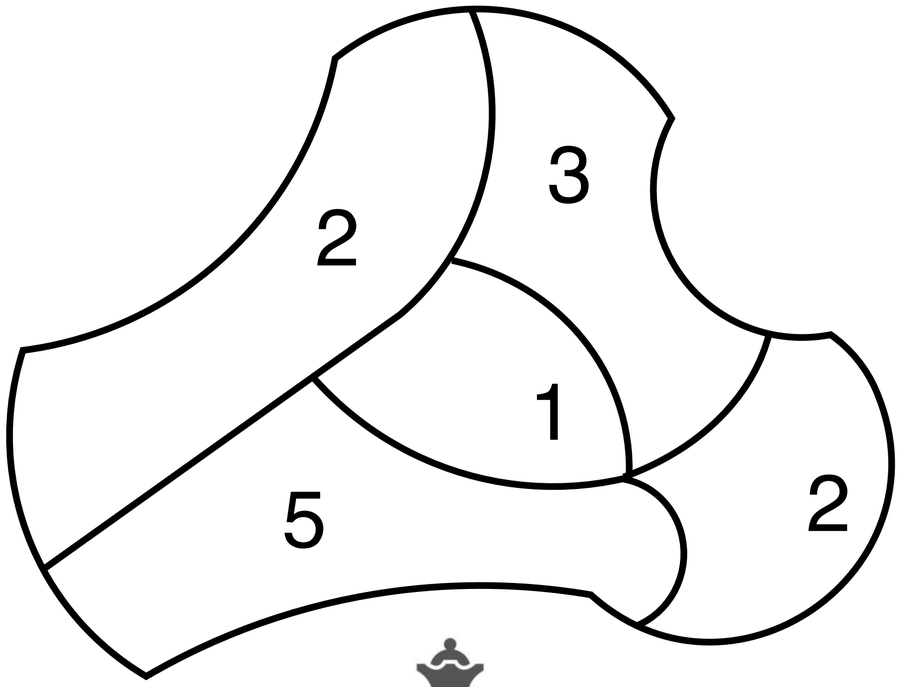
Spatial lag $\mathbf{y}_{\text{lag}}$



The spatial lag is the weighted average value of neighbors

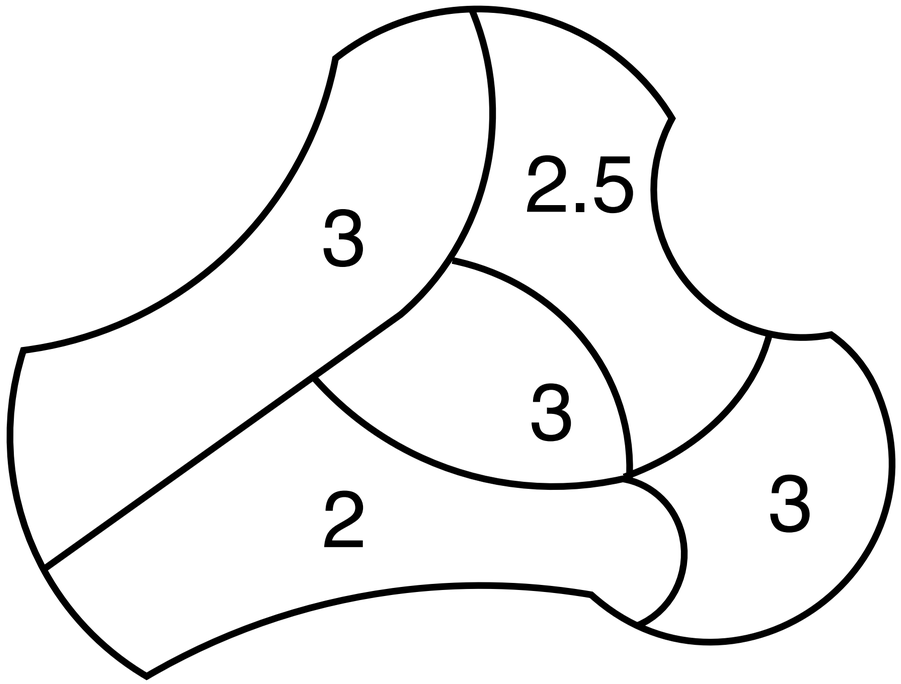
if W is standardized

Values $\mathbf{y}$



QUEEN

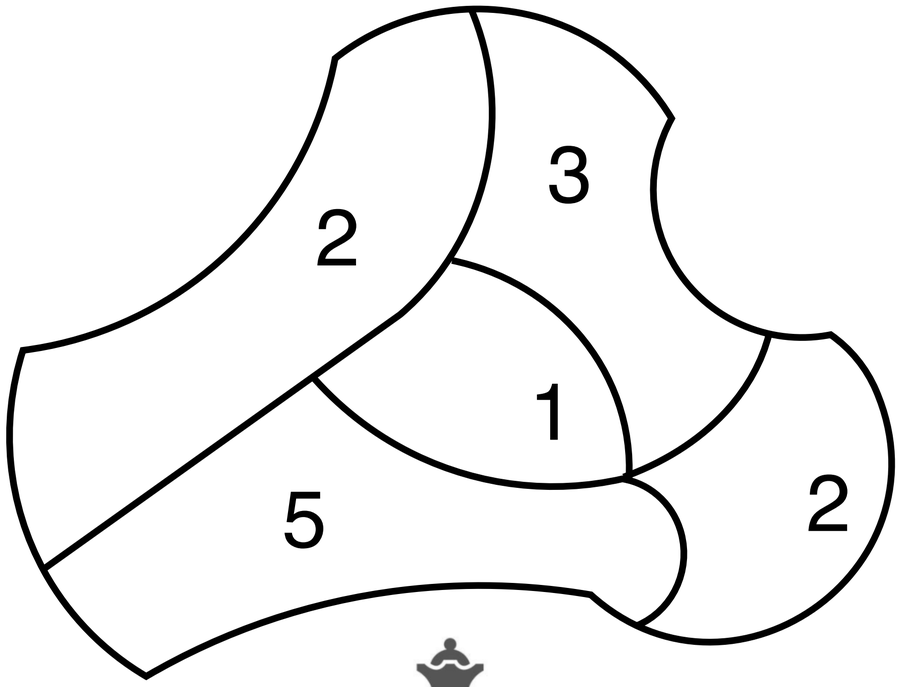
Spatial lag $\mathbf{y}_{\text{lag}}$



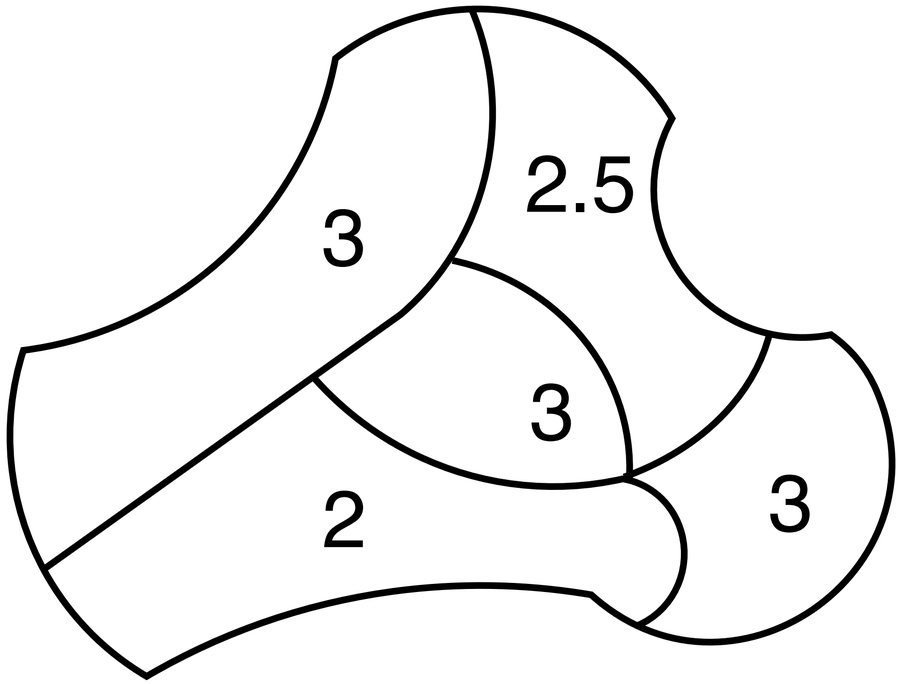
The spatial lag is the weighted average value of neighbors

if W is standardized

Values $\mathbf{y}$



Spatial lag $\mathbf{y}_{\text{lag}}$



It is a smoother: It brings all values closer to the average

The spatial lag is the sum of products of weights and values

$$y_{\text{lag},i} = w_{i1}y_i + w_{i2}y_2 + \cdots + w_{in}y_n = \sum_{j=1}^{n} w_{ij}y_j$$

$$\mathbf{y}_{\text{lag}} = \left(\sum_{j=1}^{n} w_{ij}y_j \right)_i = W\mathbf{y}$$

The spatial lag appears in many tools and models

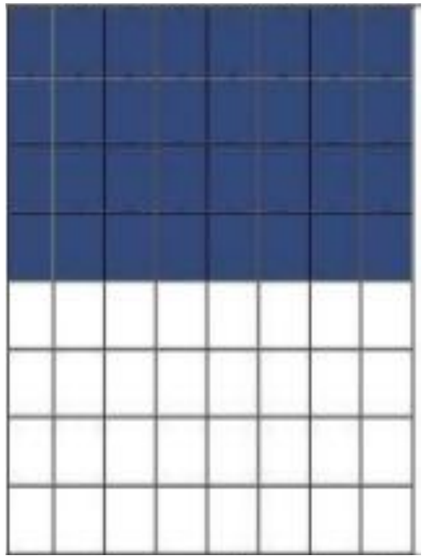
Spatial autocorrelation describes the relationship between values and locations

Spatial autocorrelation: Positive vs. Negative

Is the spatial counterpart of traditional correlation

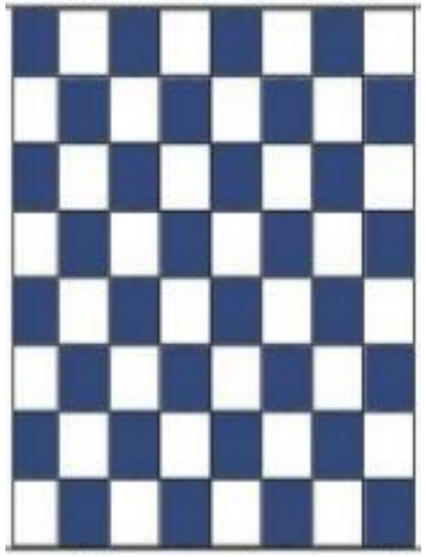
Positive

similar values are closeby



Negative

similar values are further apart

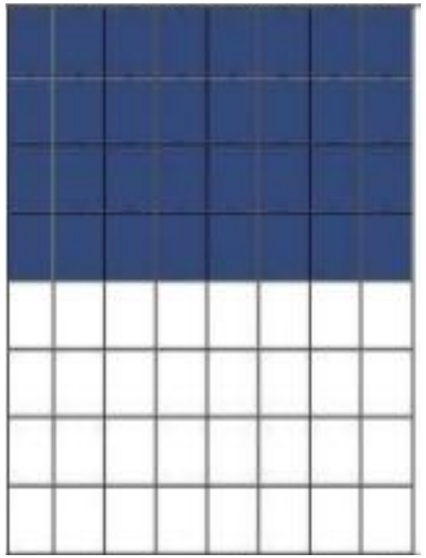


Spatial autocorrelation: Positive vs. Negative

Is the spatial counterpart of traditional correlation

Positive

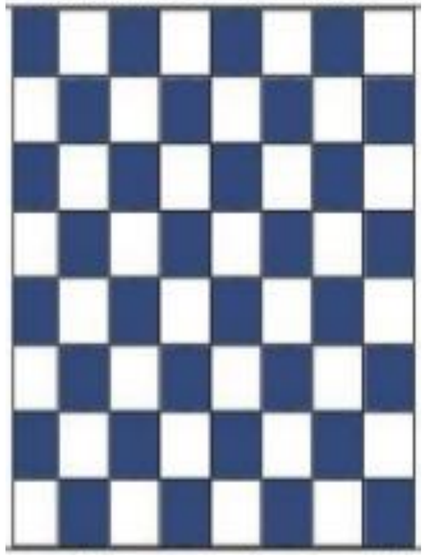
similar values are closeby



income, poverty, covid cases, vegetation, temperature,...

Negative

similar values are further apart

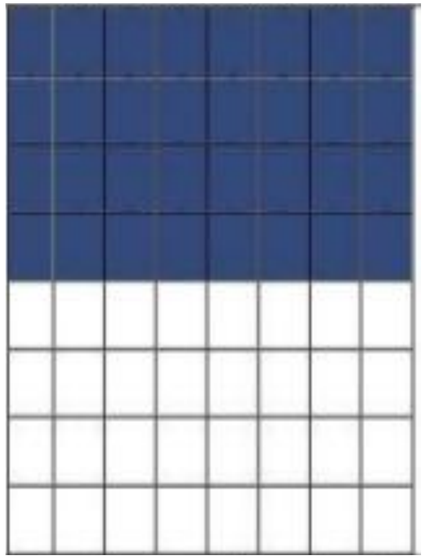


Spatial autocorrelation: Positive vs. Negative

Is the spatial counterpart of traditional correlation

Positive

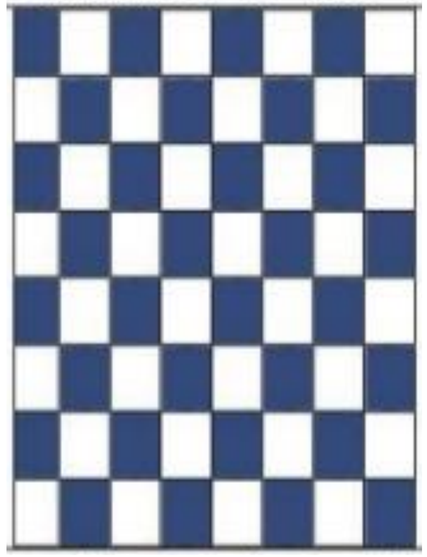
similar values are closeby



income, poverty, covid cases, vegetation, temperature,...

Negative

similar values are further apart

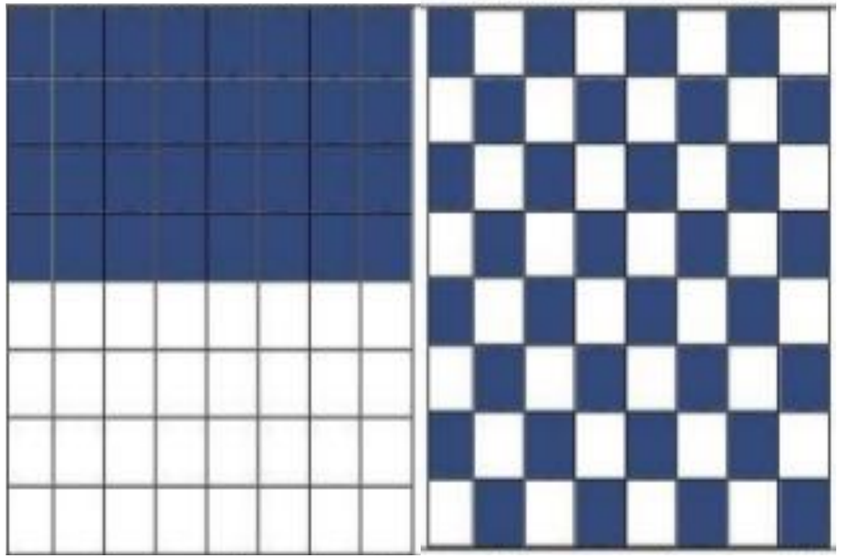


pharmacies, fire/police/metro stations, hospitals, tigers, ...

Spatial autocorrelation: Global vs. Local

Global

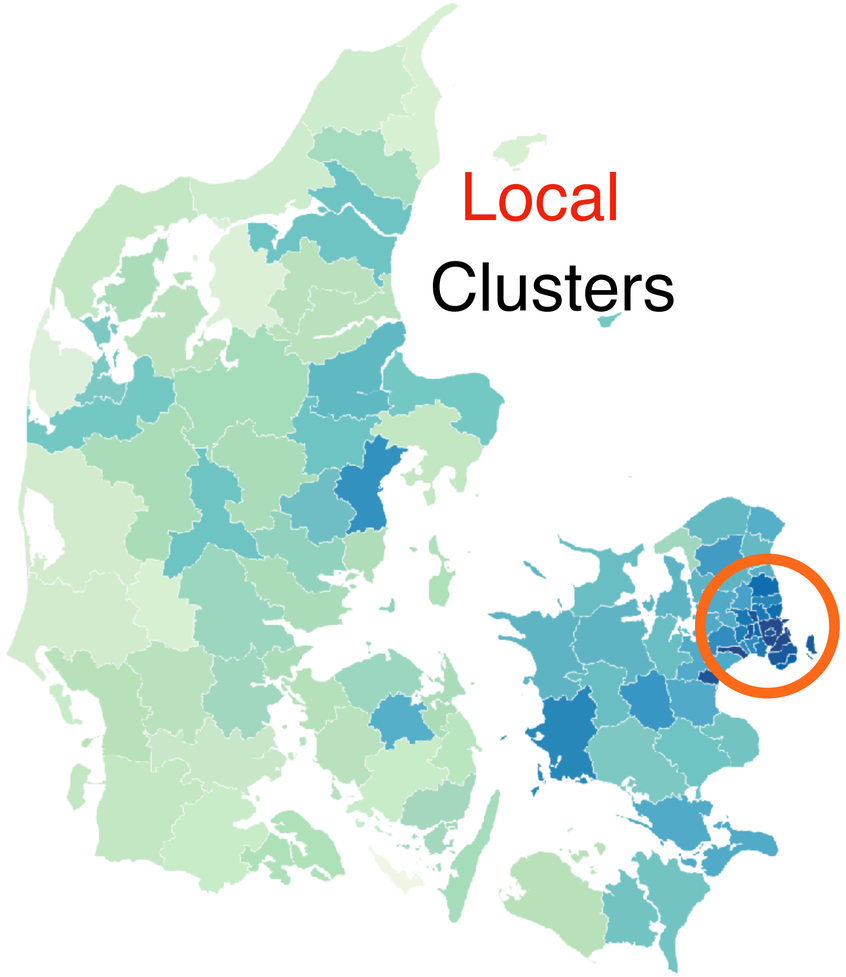
Clustering



Do values tend to be close to (dis)similar values?

Local

Clusters

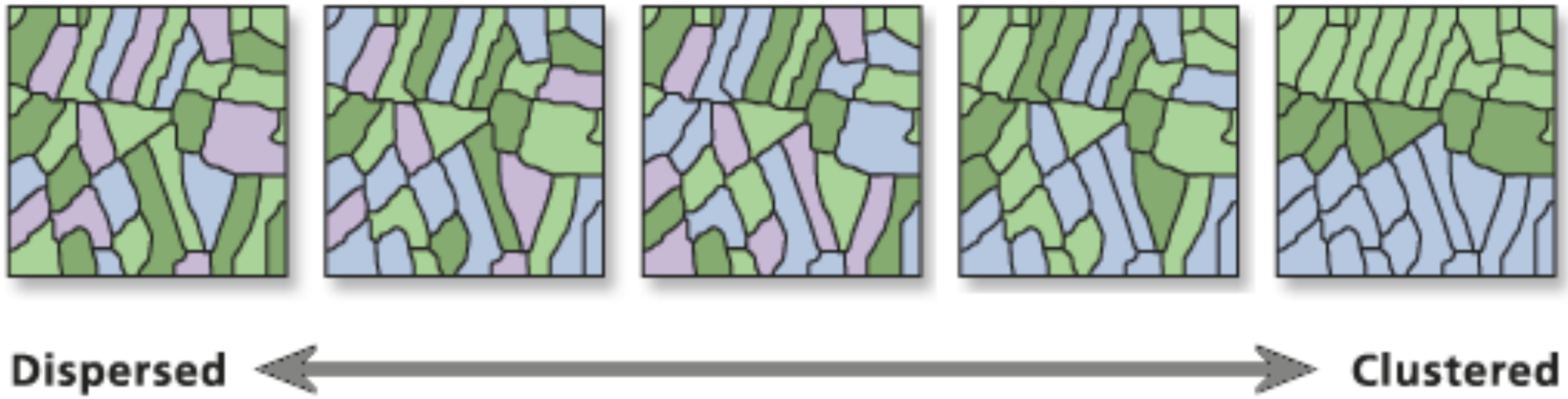


Are there areas with an extraordinary concentration of (dis)similar values?

Global Spatial Autocorrelation

Global spatial autocorrelation: Moran's I

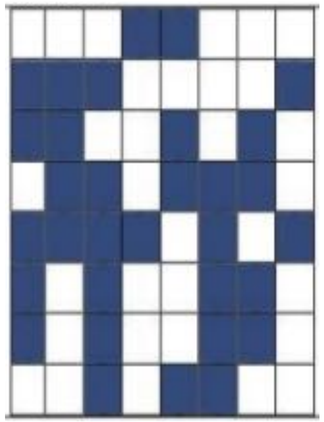
Moran's I measures the average correlation between the value of a variable at one location and the value at nearby locations.



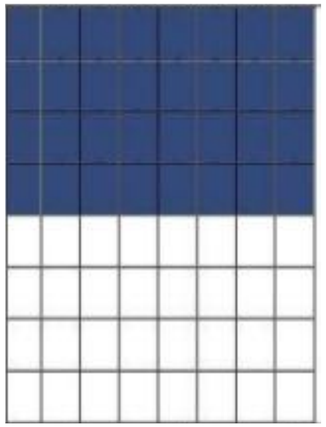
Global spatial autocorrelation: Moran's I

How likely is it to obtain a map like observed from a purely random pattern?

$I = -1$



$I \approx 0$



$I \approx 1$

Global spatial autocorrelation: Moran's I

Moran's I measures the average correlation between the value of a variable at one location and the value at nearby locations.

y_i

$$y_{\text{lag},i} = \sum_{j=1}^{n} w_{ij} y_j$$

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - \bar{Y}) (y_j - \bar{Y})}{\sum_i (y_i - \bar{Y})^2}$$

Global spatial autocorrelation: Moran's I

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - \bar{Y}) (y_j - \bar{Y})}{\sum_i (y_i - \bar{Y})^2}$$

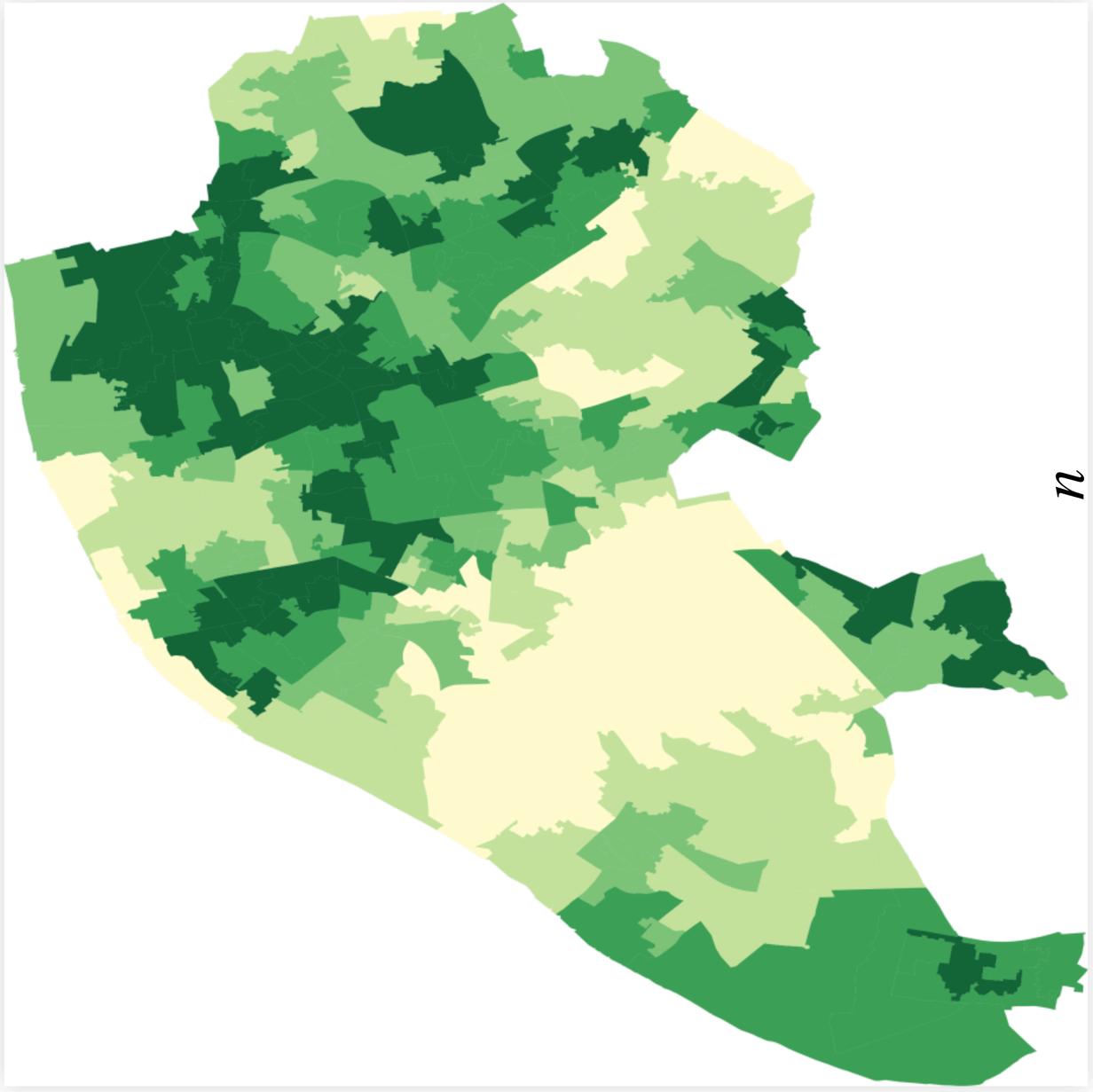
Standardized Moran's I:

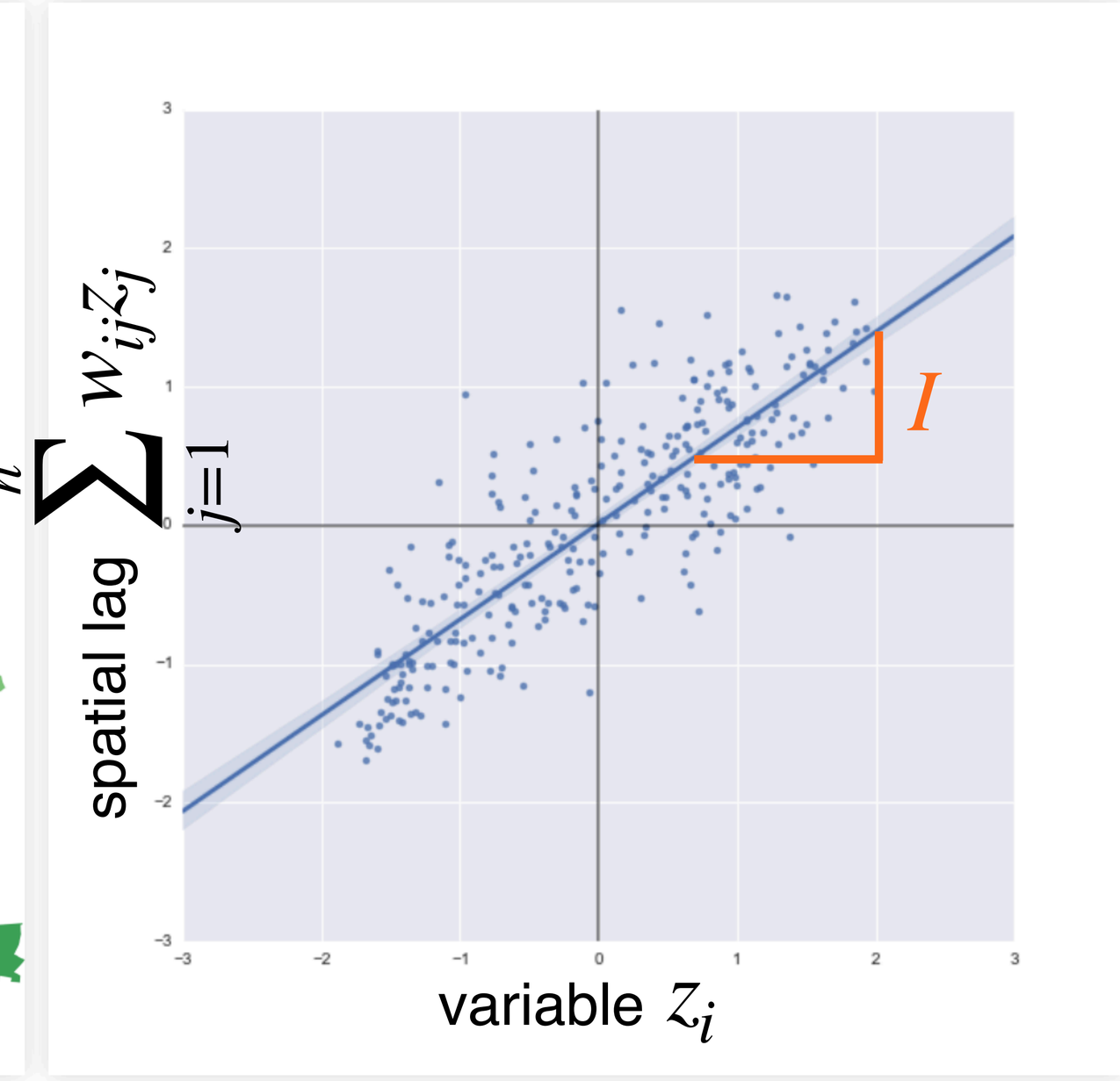
If we call $z_i = \left(\frac{y_i - \bar{y}}{s_y} \right)$, then: $$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

Row standardized Moran's I:

$$I = \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

Moran's I is the regression slope in the Moran plot





Local Spatial Autocorrelation

Local spatial autocorrelation is about local clusters

Cluster = Portion of a map where values are correlated in a particularly strong or specific way

Local spatial autocorrelation is about local clusters

Cluster = Portion of a map where values are correlated in a particularly strong or specific way

Positive

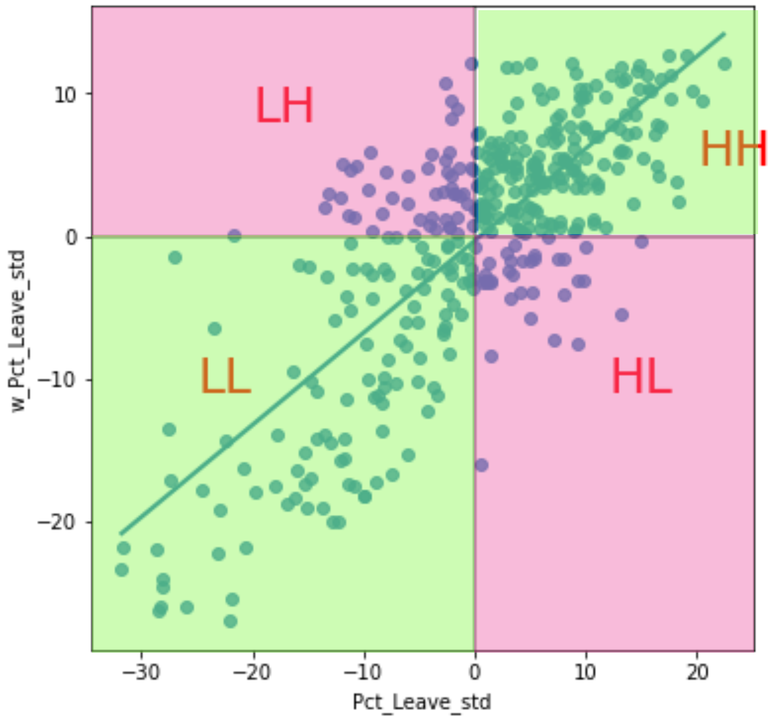
High-High: Hotspot

Low-Low: Coldspot

Negative

High-Low: Spatial outlier

Low-High: Spatial outlier (donut)



https://geographicdata.science/book/notebooks/07_local_autocorrelation.html

LISA: Local Indicators of Spatial Association

Like a local Moran's I

How much is each object's relation with its neighbors different from the relations of other objects and their neighbors on the map?

Anselin, Geographical analysis 27.2 (1995)

LISA: Local Indicators of Spatial Association

Like a local Moran's I

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j \qquad m_2 = \frac{\sum_i z_i^2}{n} \qquad z_i = y_i - \bar{y}$$

Anselin, Geographical analysis 27.2 (1995)

LISA: Local Indicators of Spatial Association

Like a local Moran's I

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j \qquad m_2 = \frac{\sum_i z_i^2}{n} \qquad z_i = y_i - \bar{y}$$

If W is row-standardized, then:

$$\sum_i I_i = \gamma I$$

$$\gamma = \sum_i \sum_j w_{ij} \quad = \text{scale factor}$$

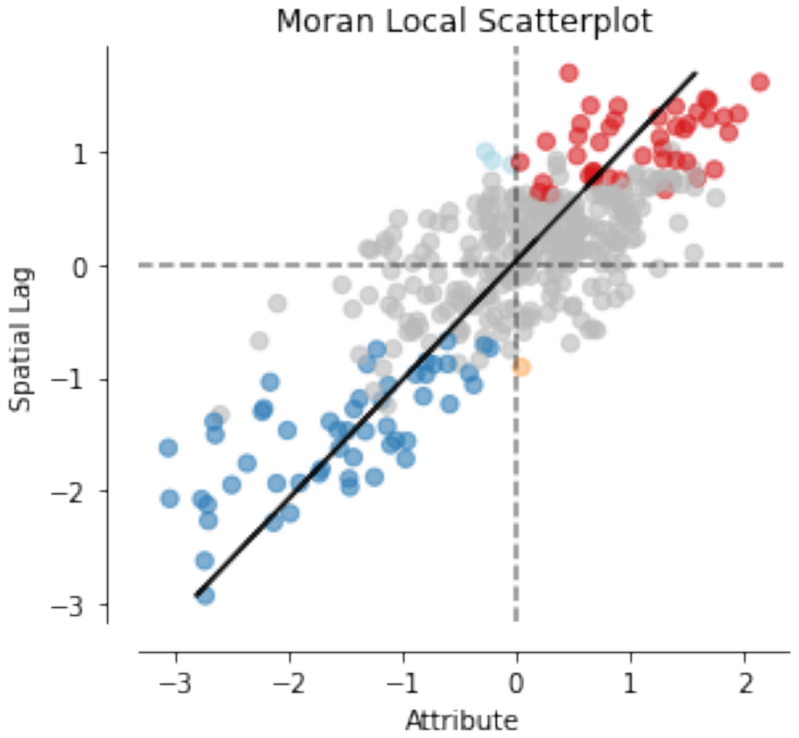
$I =$ global indicator of autocorrelation

Sum of local indicators is proportional to global indicator

Anselin, Geographical analysis 27.2 (1995)

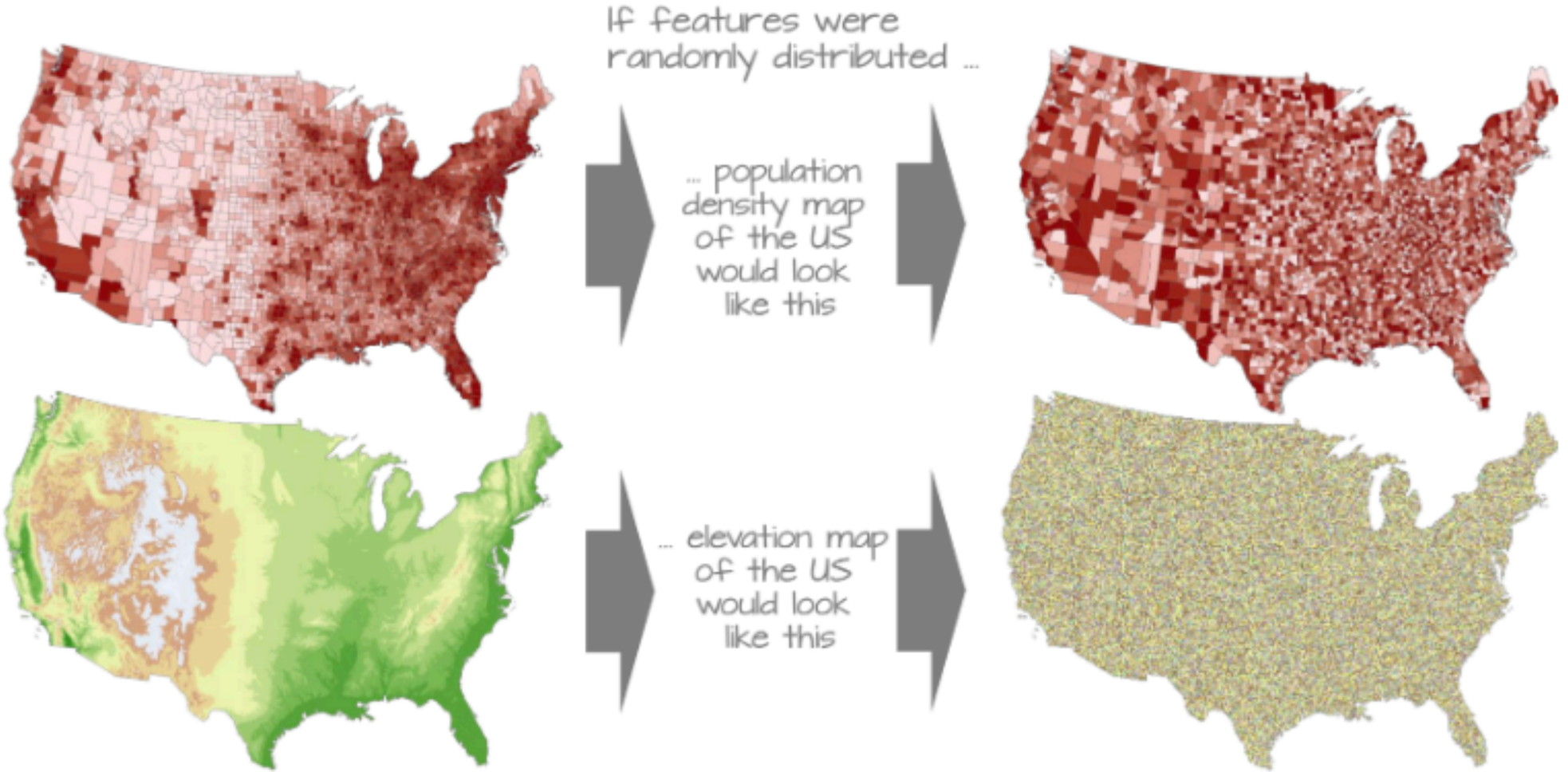
We need to test whether the pattern is significant

To what extent is each object's relation with its neighbors significantly different from the relations of other objects and their neighbors on the map?



We need to test whether the pattern is significant

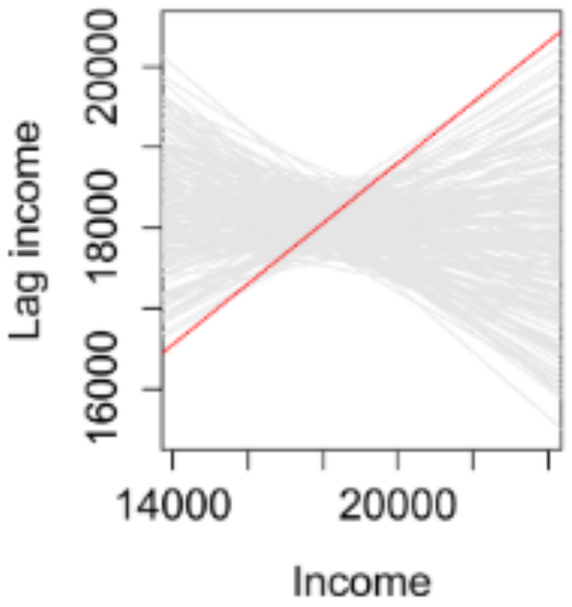
Null hypothesis is a random pattern

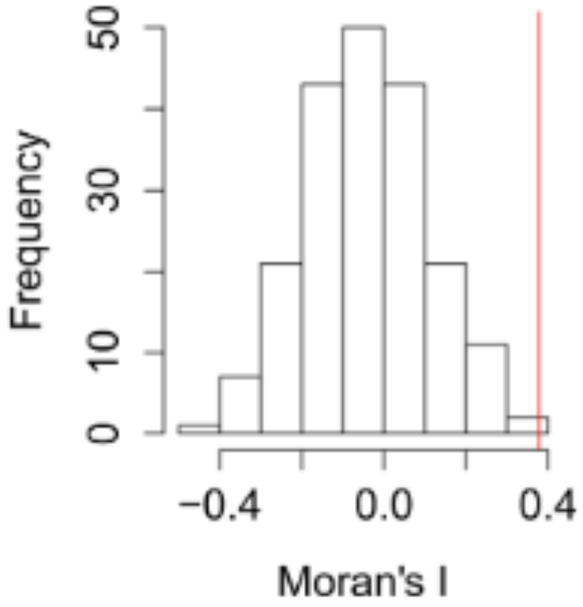


https://mgimond.github.io/Spatial/spatial-autocorrelation.html

We need to test whether the pattern is significant

Null hypothesis is a random pattern





Randomize many times, measure p-value

https://mgimond.github.io/Spatial/spatial-autocorrelation.html

Applications of spatial autocorrelation

Any analysis where we need to understand if a *spatial* process/relation is happening

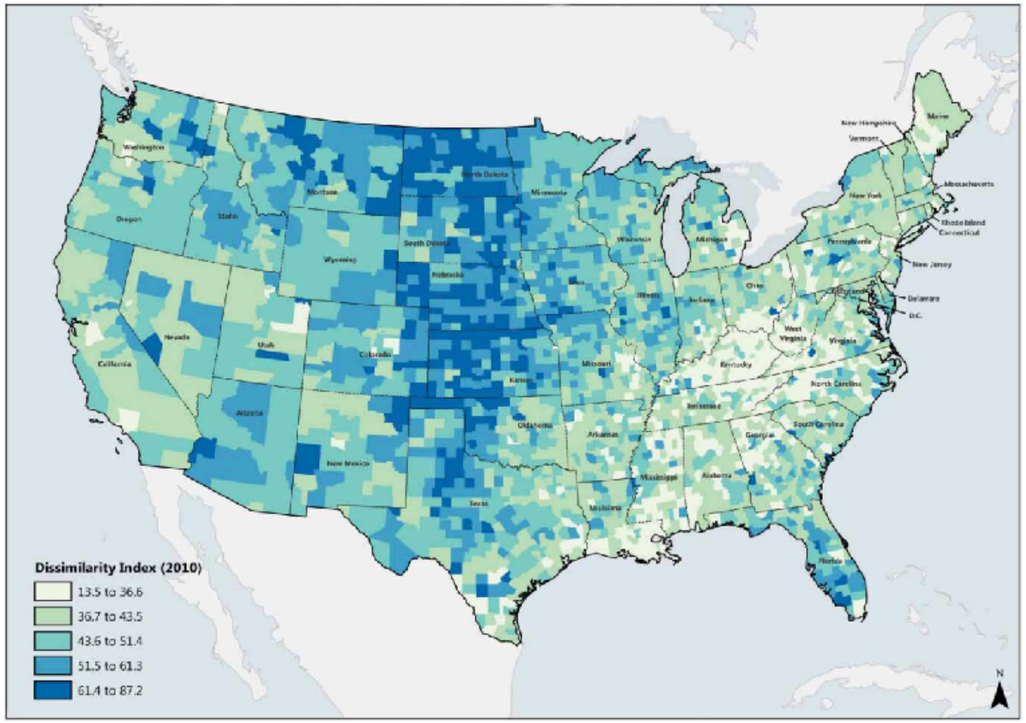


Figure 1. Residential segregation between older and younger adults within counties.

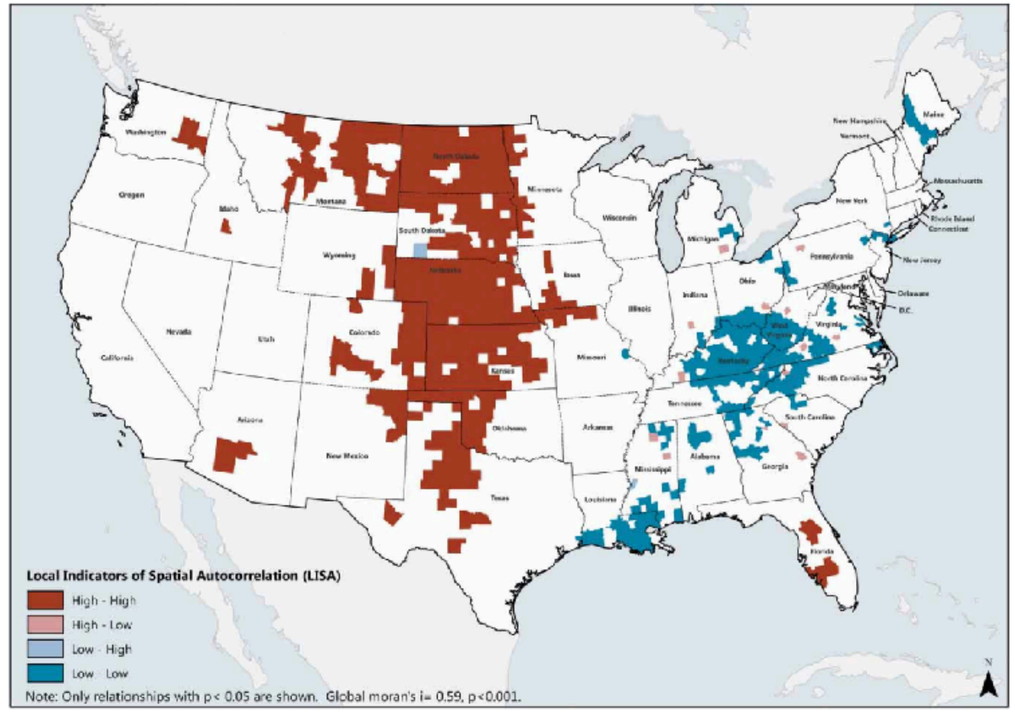


Figure 2. Clusters of age segregation: dissimilarity index of blocks within counties.

Applications of spatial autocorrelation

Any analysis where we need to understand if places are significantly different

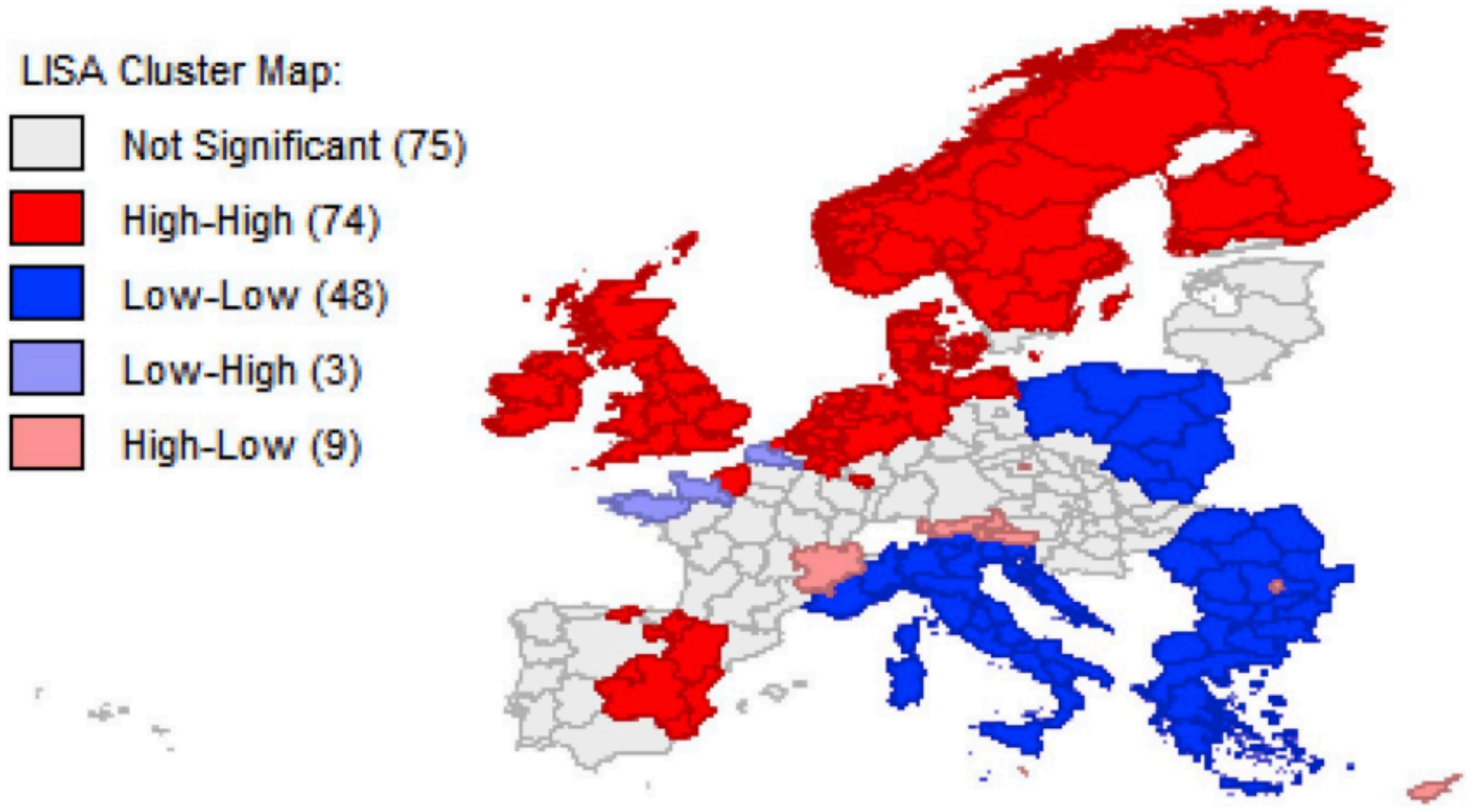


Fig. 12. 2016 proportion of population with on-the-go internet access (n = 209, Moran's I = 0.682).

Jupyter

part2/part2spatialstatistics.ipynb

github.com/NERDSITU/gdstutorial

Sources and further materials for Part 2

Geographic Data Science with Python

https://geographicdata.science/book/notebooks/05_choropleth.html

https://darribas.org/gds_course/content/bD/concepts_D.html

https://geographicdata.science/book/notebooks/06_spatial_autocorrelation.html

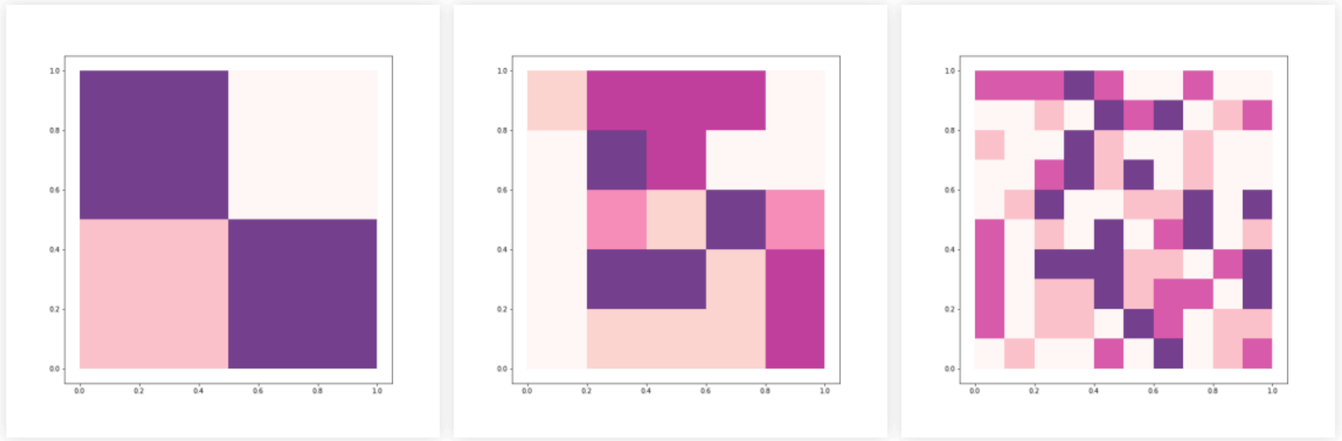
https://geographicdata.science/book/notebooks/07_local_autocorrelation.html

https://darribas.org/gds_course/content/bF/concepts_F.html

https://mgimond.github.io/Spatial/spatial-autocorrelation.html

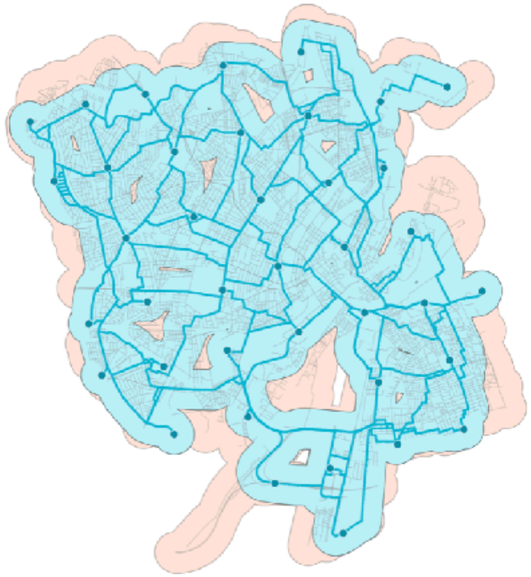
Take home messages for Part 1 & 2

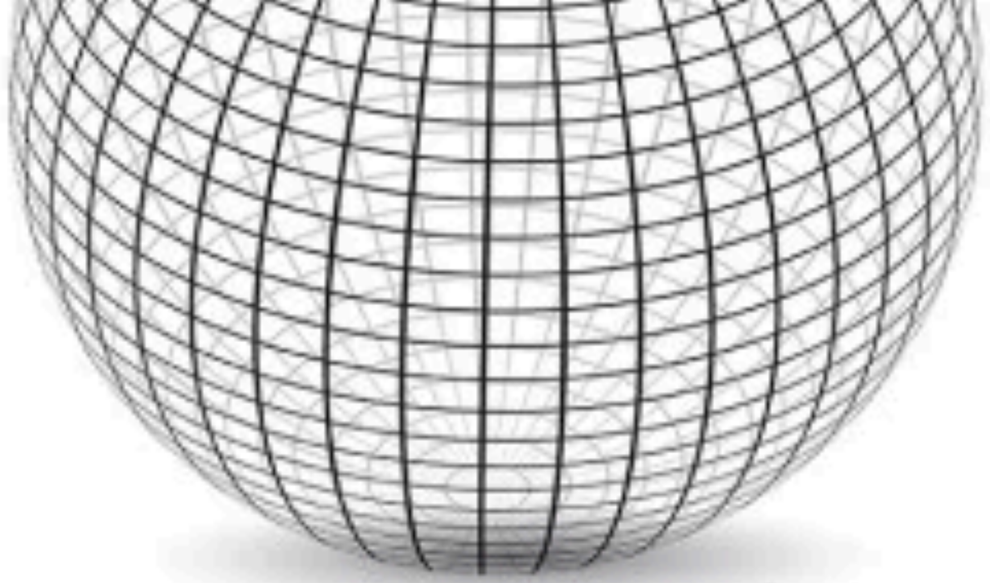
Mind the pitfalls: CRS, MAUP, classification

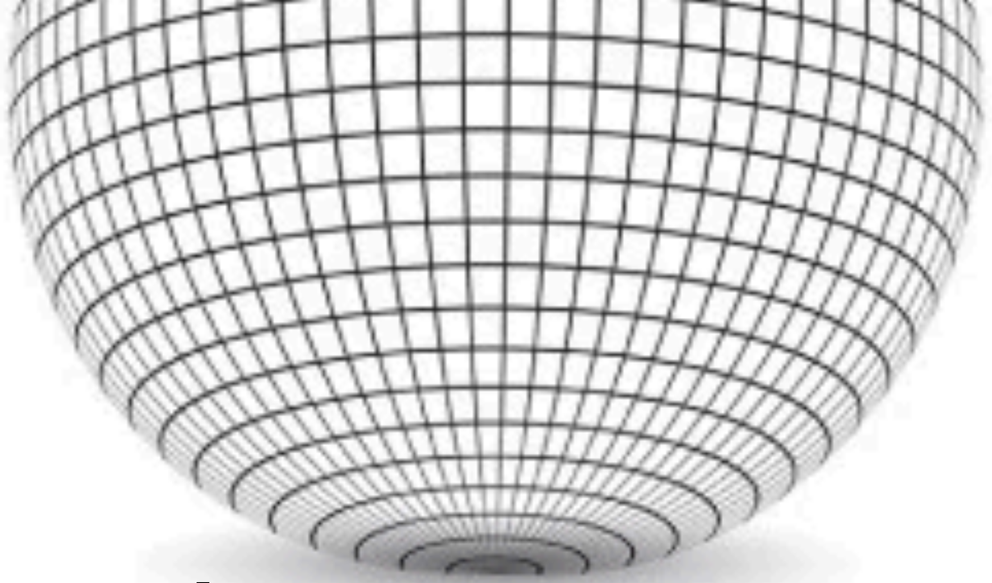


You can create *arbitrary* interpretations of the same data!

Next Parts 3 & 4: OpenStreetMap, Spatial Networks







30 min break

See you at 11:00