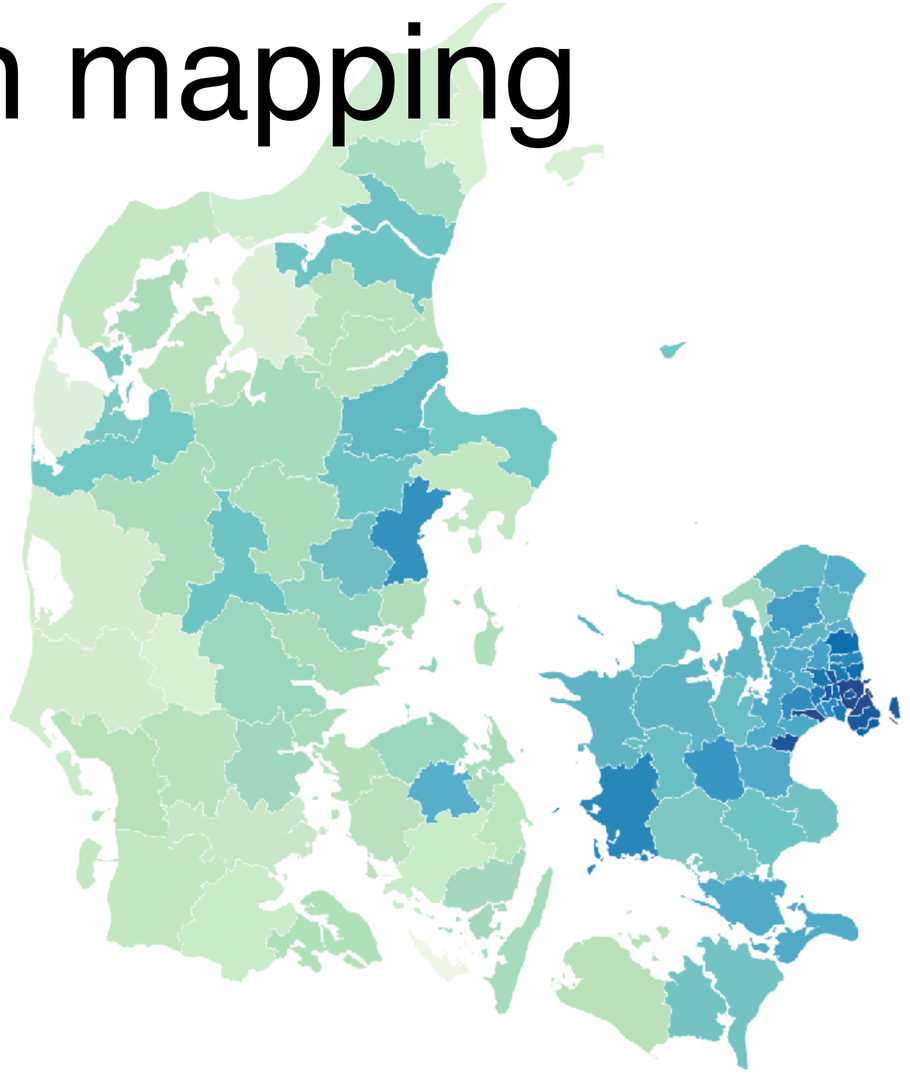
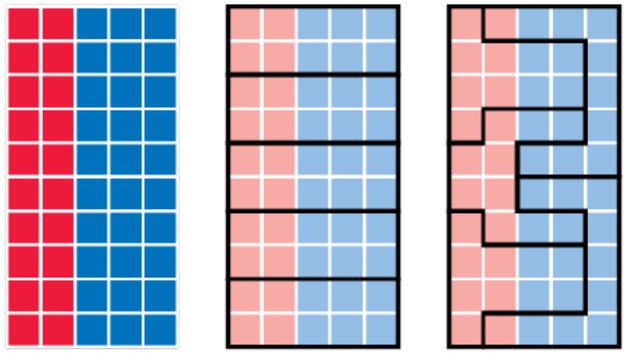
Lecture 3: Choropleth mapping

Instructor: Michael Szell
Feb 17, 2022



Today you will learn about MAUP and Choropleths

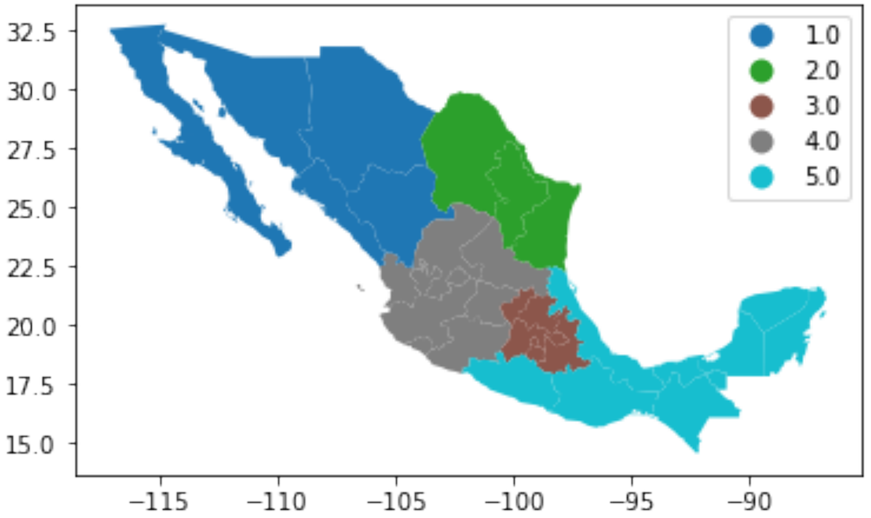
Biases from spatial aggregation (MAUP)



Classification schemes

$$c_j < y_i \leq c_{j+1} \; \forall y_i \in C_j$$

Choropleths in Python



The most common steps in Data Preprocessing are:

Aggregation

Sampling

Dimensionality reduction

Discretization / Classification

Variable transformation

Aggregation = Combining objects into a single one

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

NULL	Non-Freshman	3.375

Aggregation = Combining objects into a single one

Examples:

GPS coordinate → Zip Code → City → Country

Second → Minute → Hour → Day → Week → Month → Year

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

Disadvantages: Loss of details, introducing biases

Choropleth maps visualize geospatial data with colors

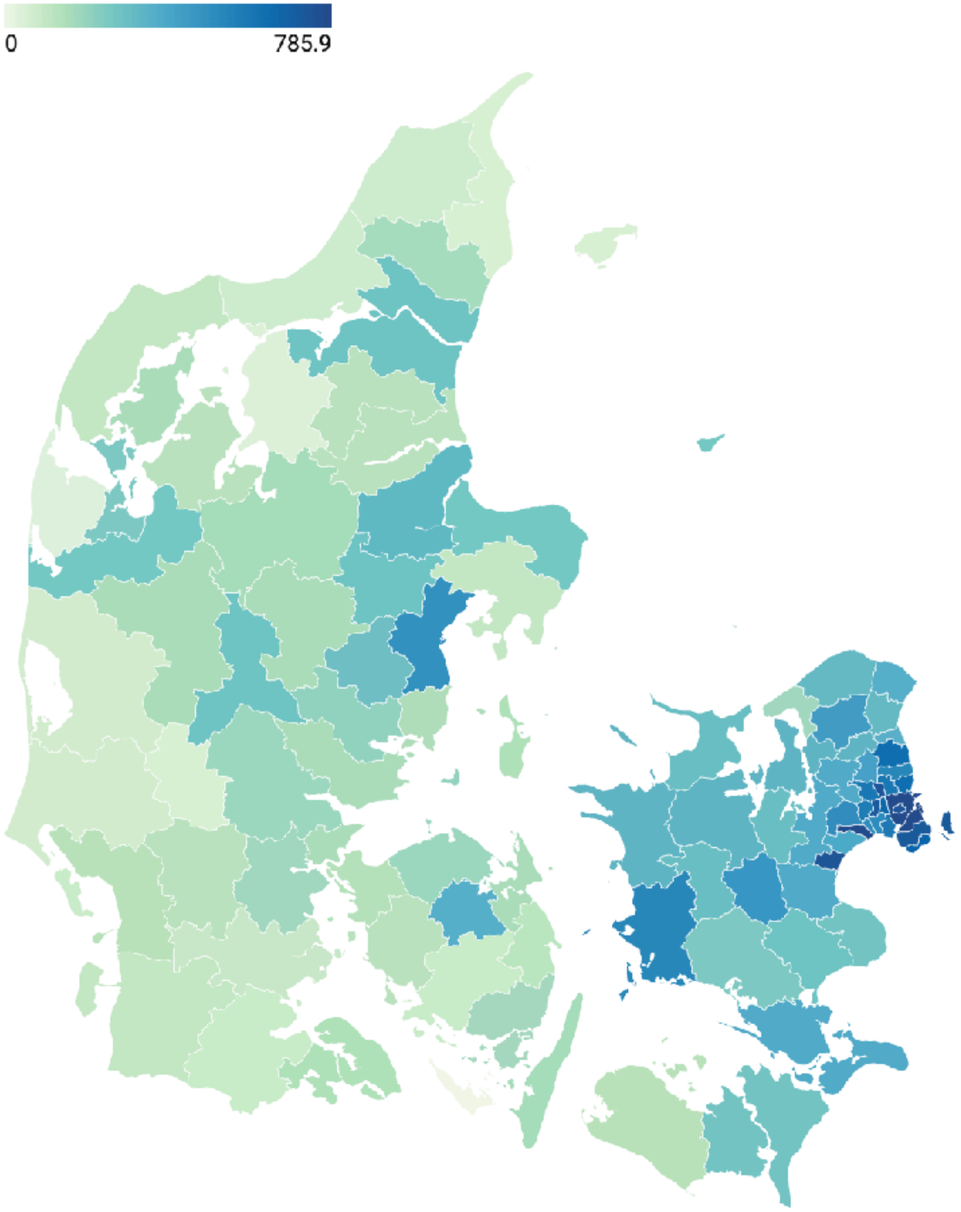
Choro = region
pleth = multiple

Combine: Polygons + Attribute values

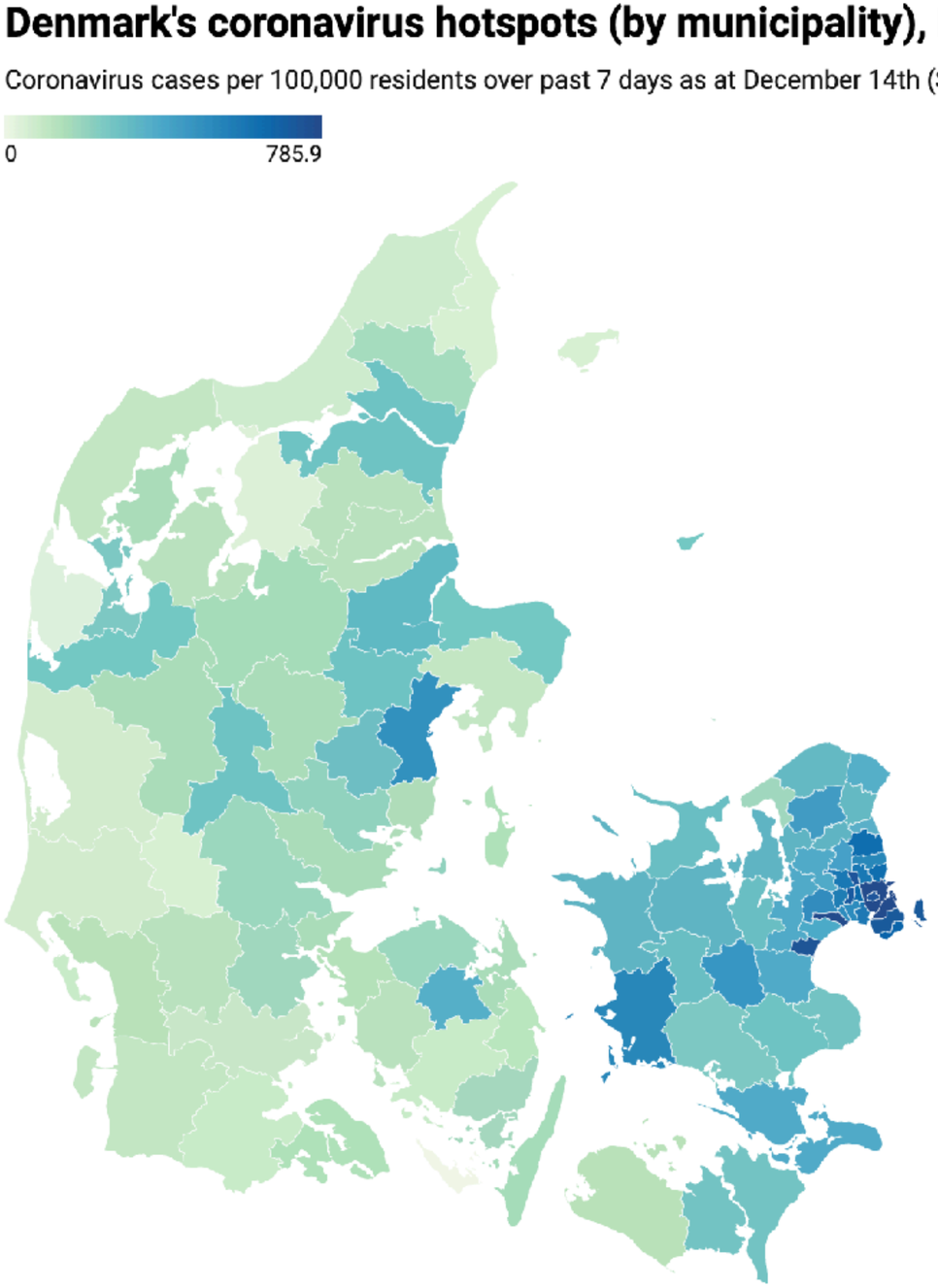
Values are often grouped into classes to reduce complexity

Denmark's coronavirus hotspots (by municipality), December 14th

Coronavirus cases per 100,000 residents over past 7 days as at December 14th (Source: SSI)



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



3 choices:

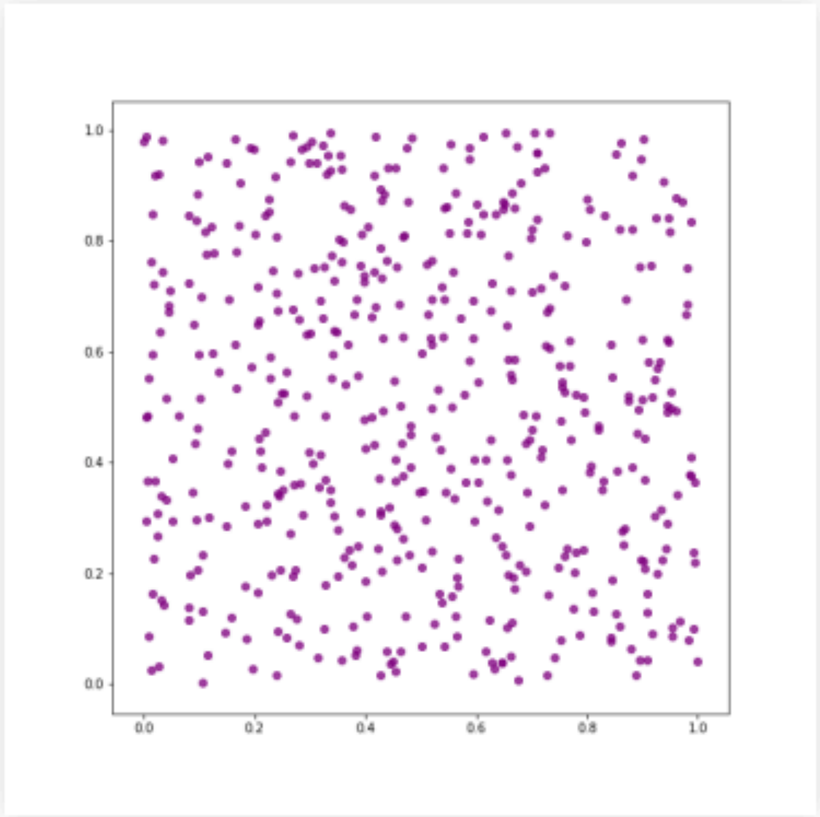
1) spatial units
2) colors
3) classes

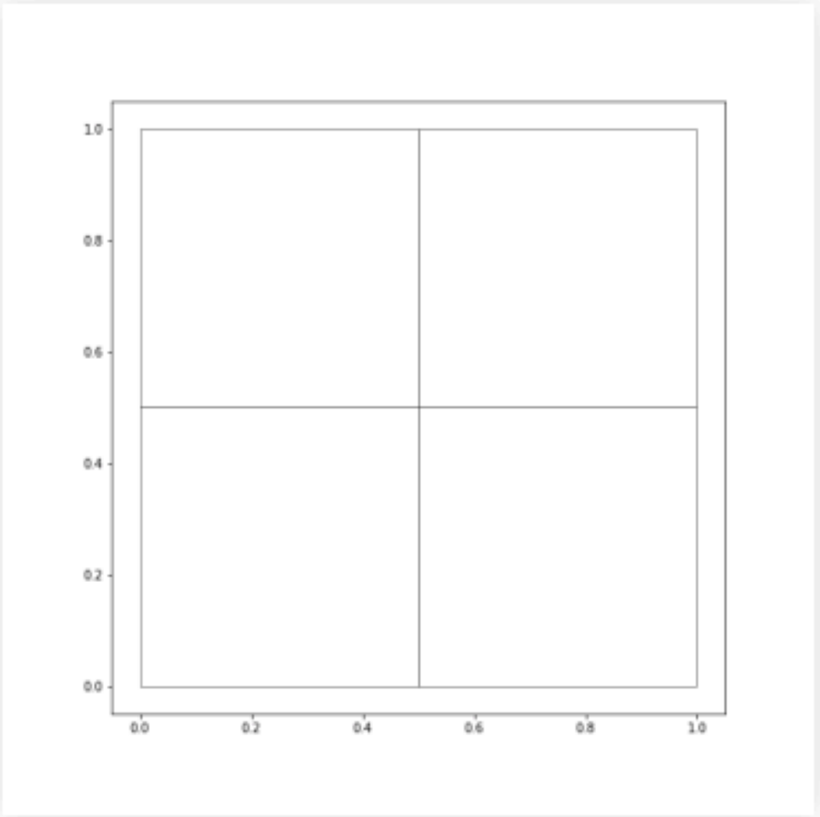
https://www.thelocal.dk/20201028/maps-where-are-denmarks-coronavirus-hotspots/

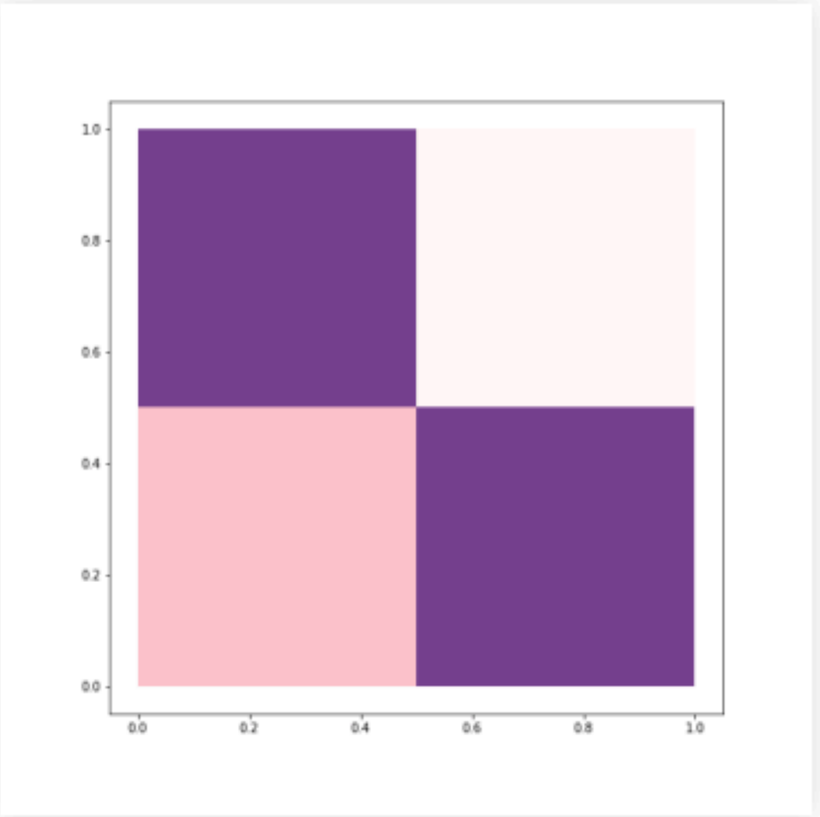
Population density (people aggregated in space)

Group experiment

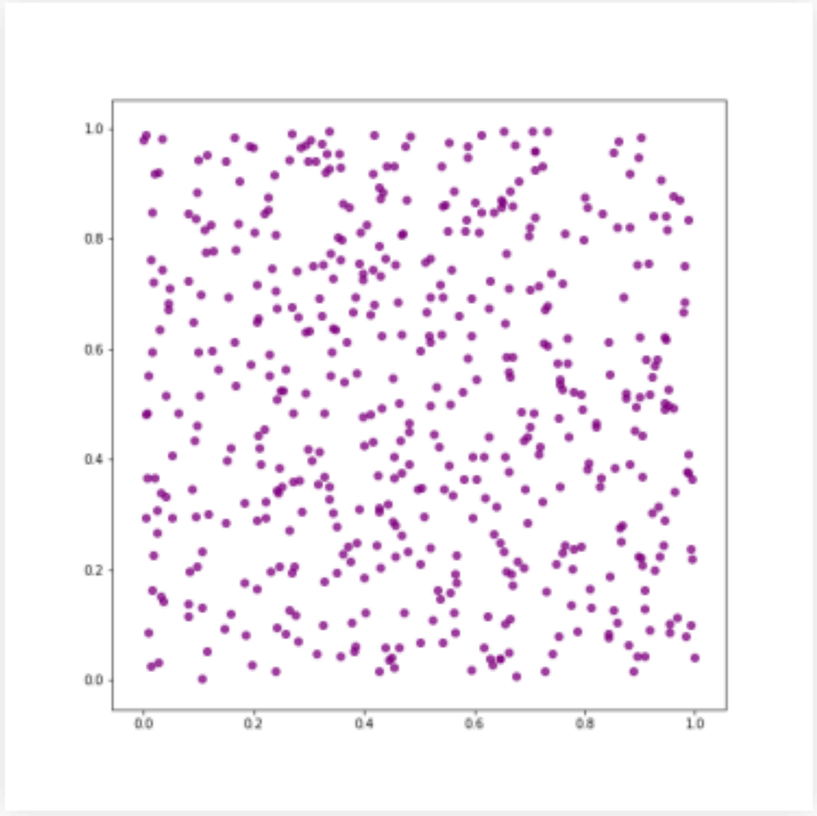
Each of 3 groups describes their picture in 1 sentence

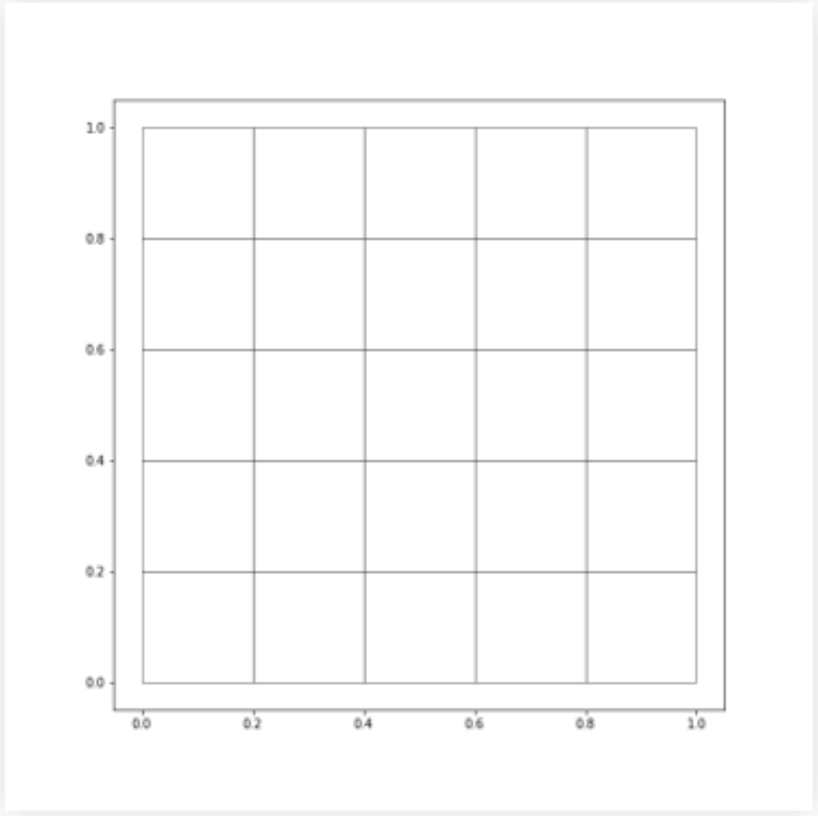


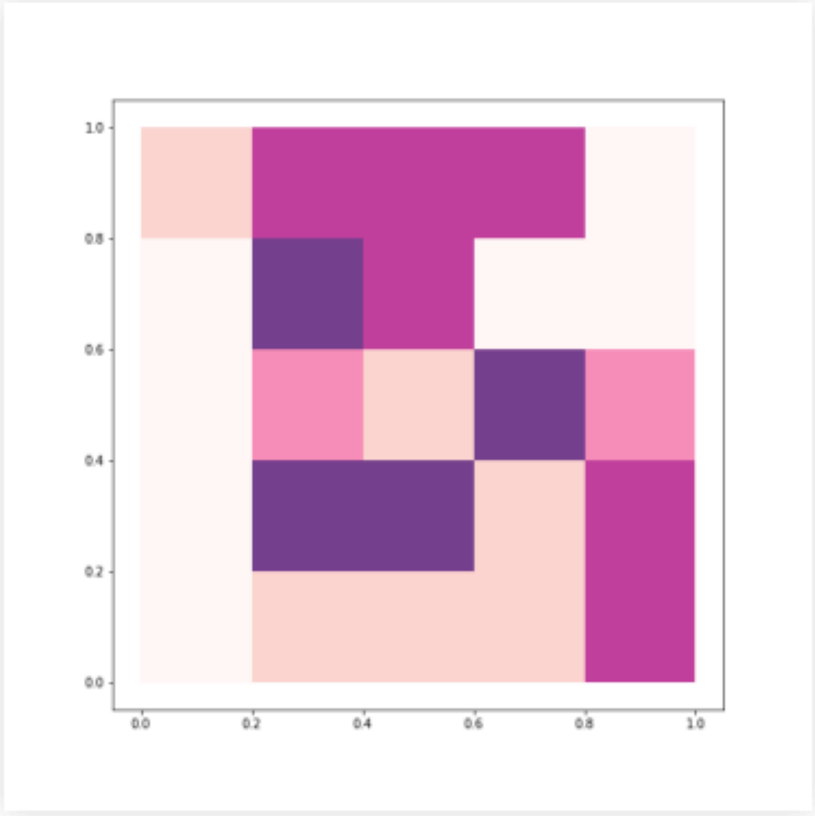




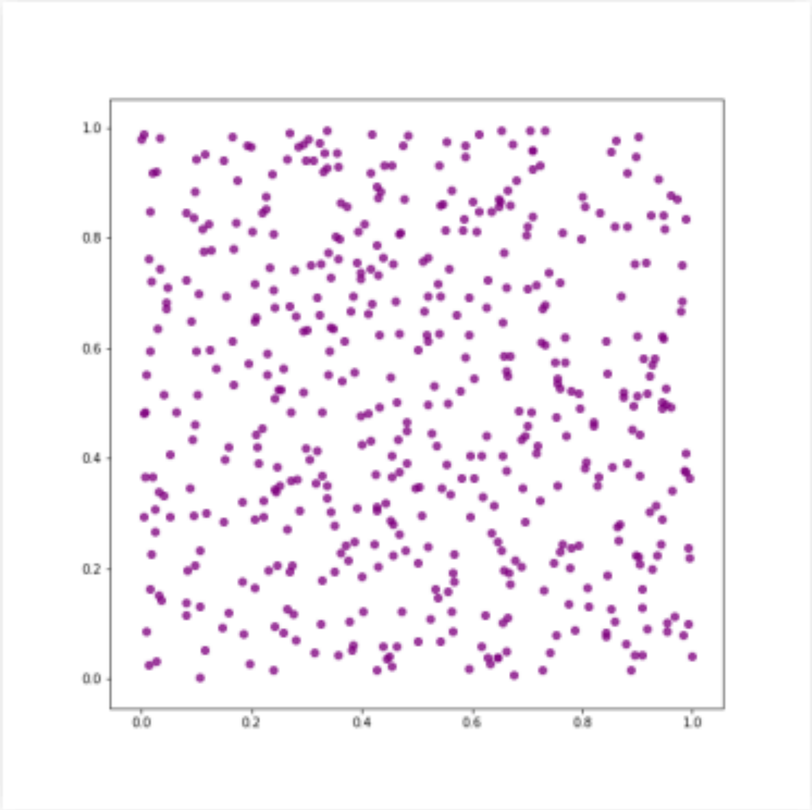
https://gist.github.com/darribas/8b5a7b93d4085223f1c5

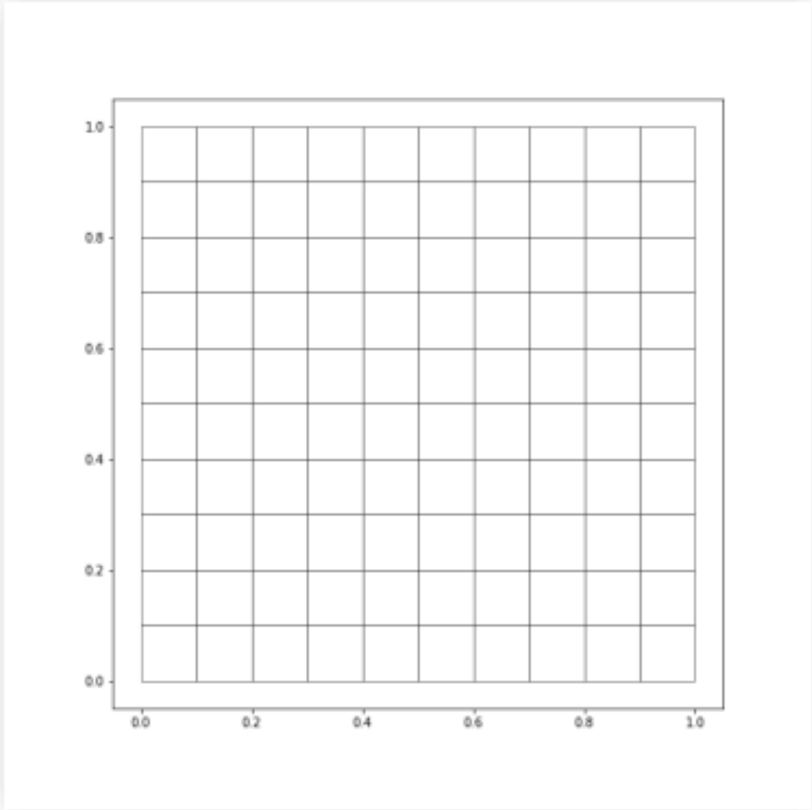


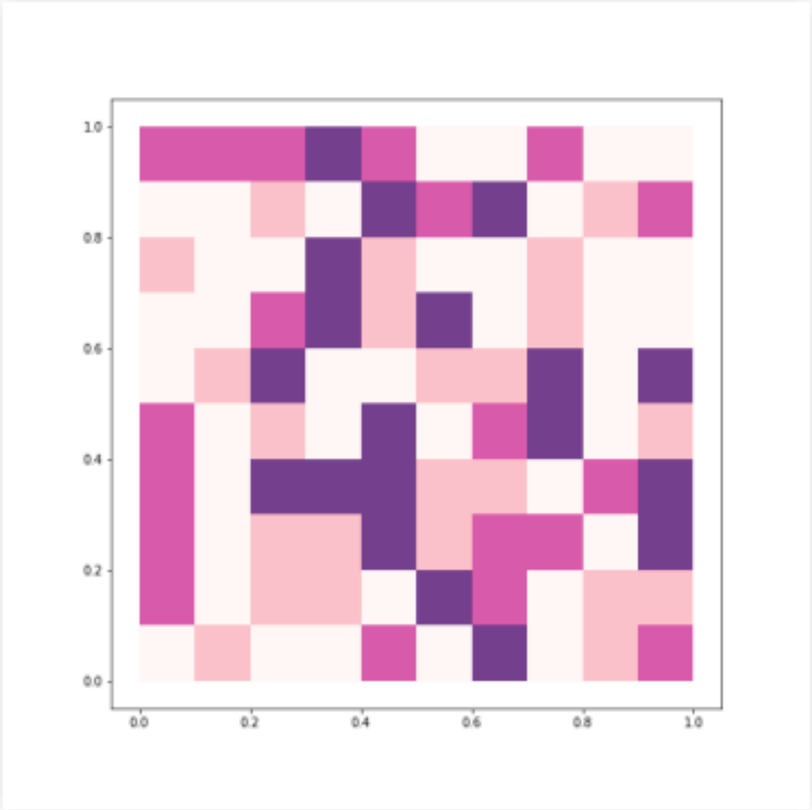




https://gist.github.com/darribas/8b5a7b93d4085223f1c5

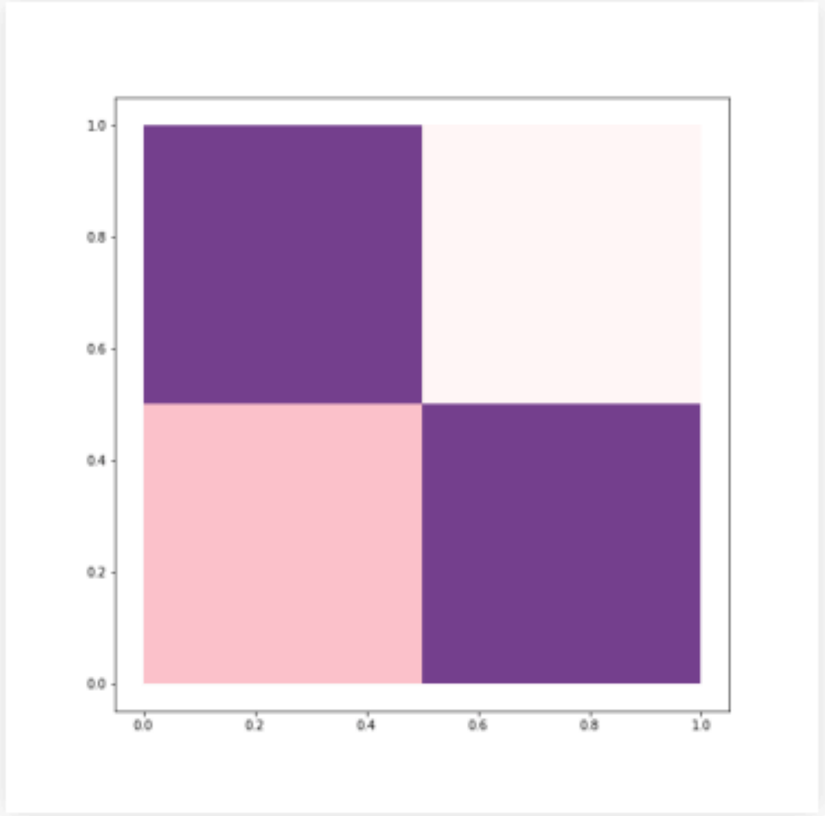


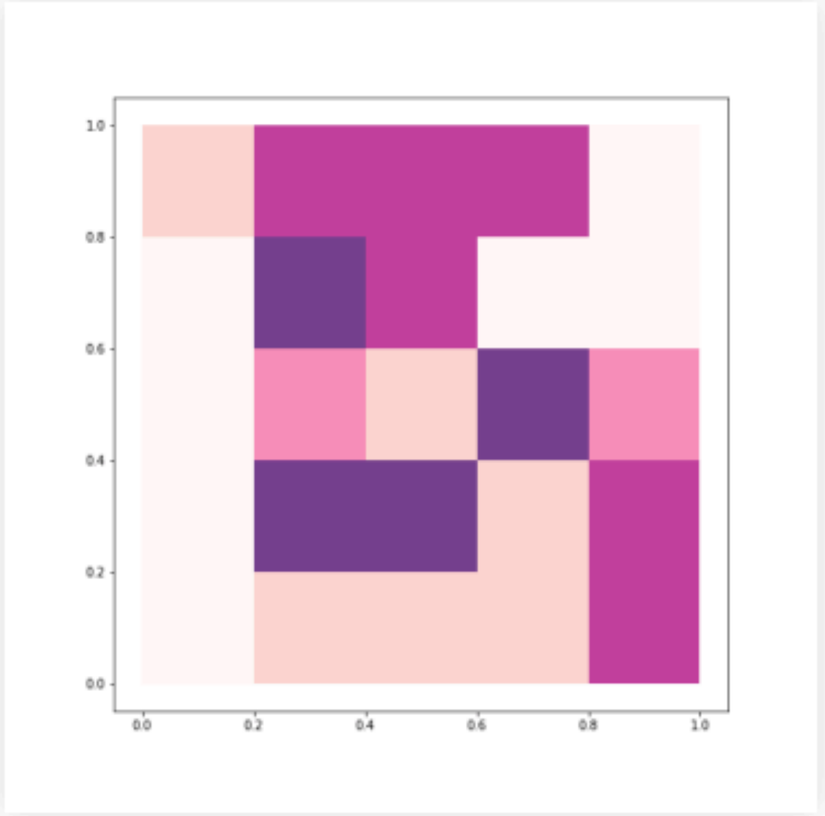


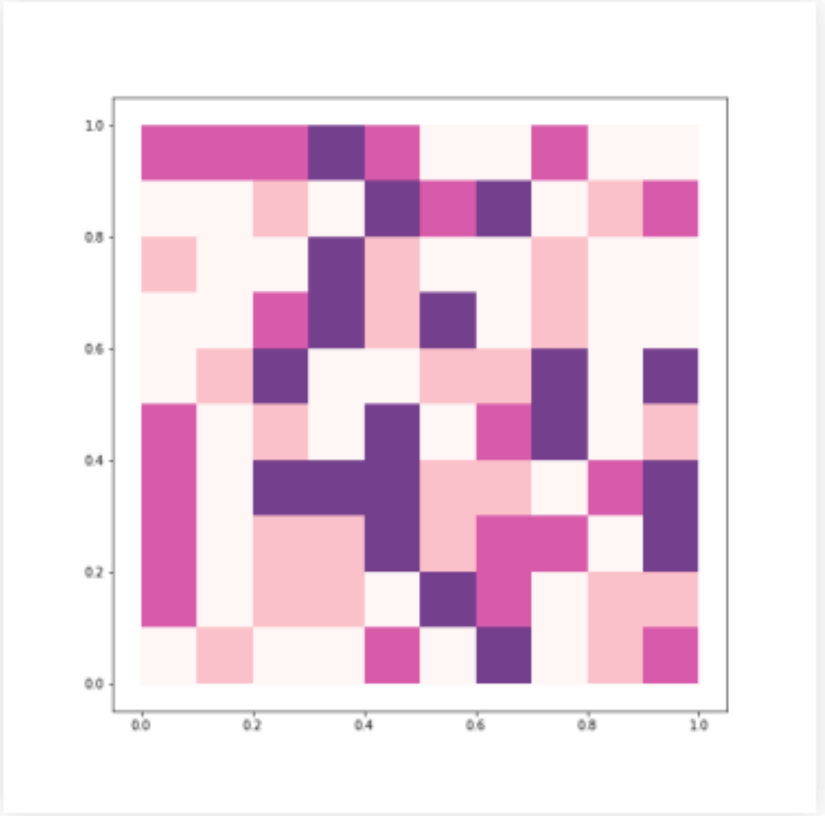


https://gist.github.com/darribas/8b5a7b93d4085223f1c5

A common source of bias in spatial aggregation: MAUP





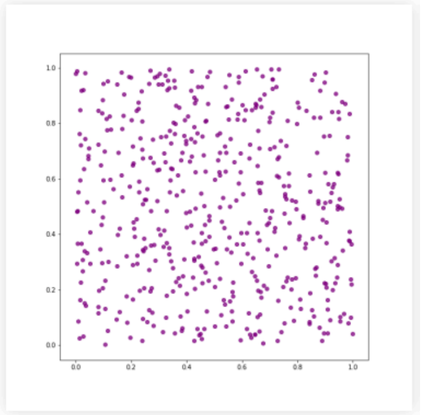


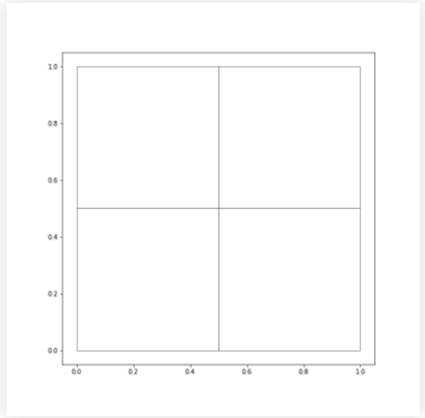
https://gist.github.com/darribas/8b5a7b93d4085223f1c5

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



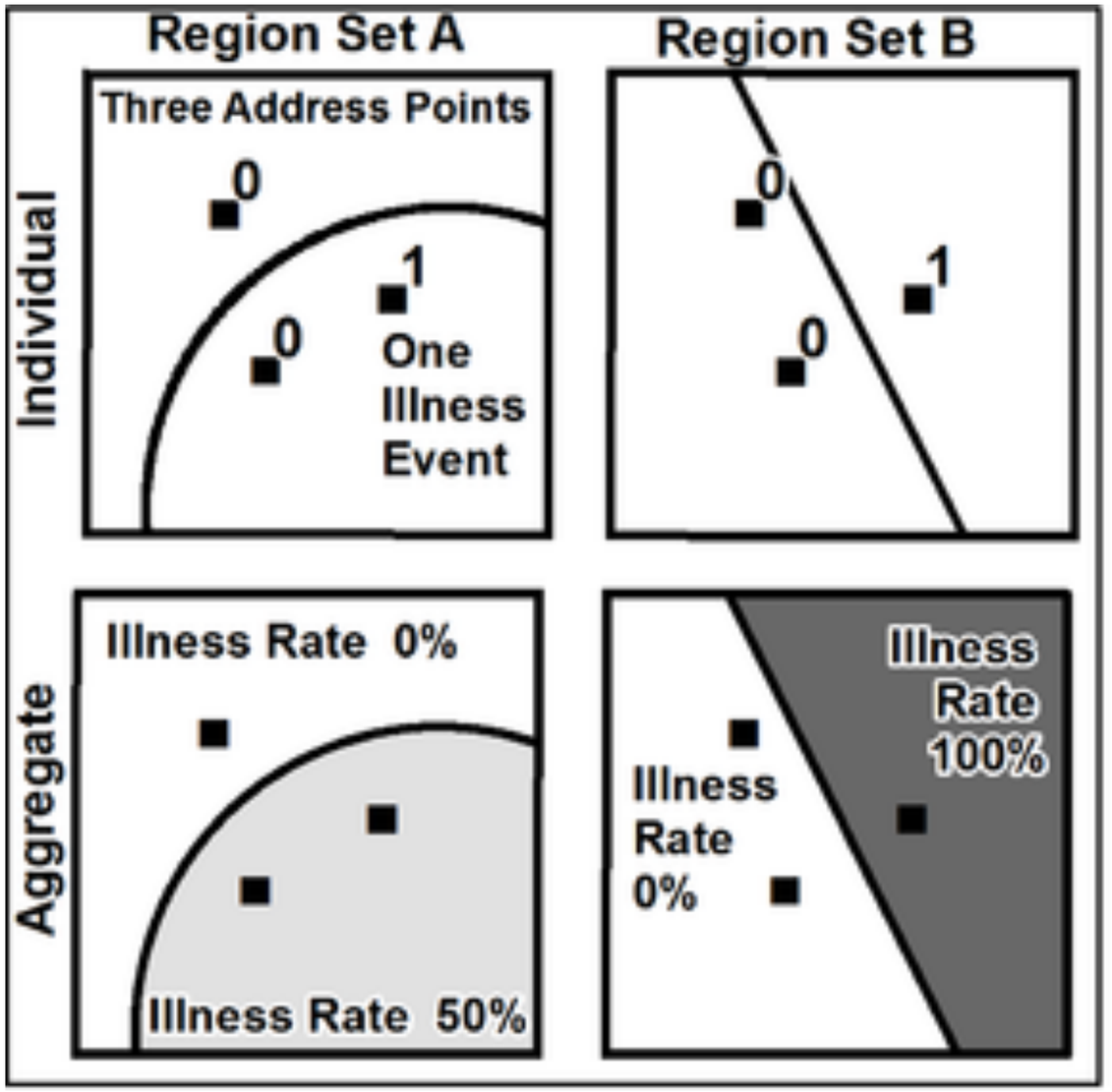


Individuals, shops,..

Districts, regions, resolution,..

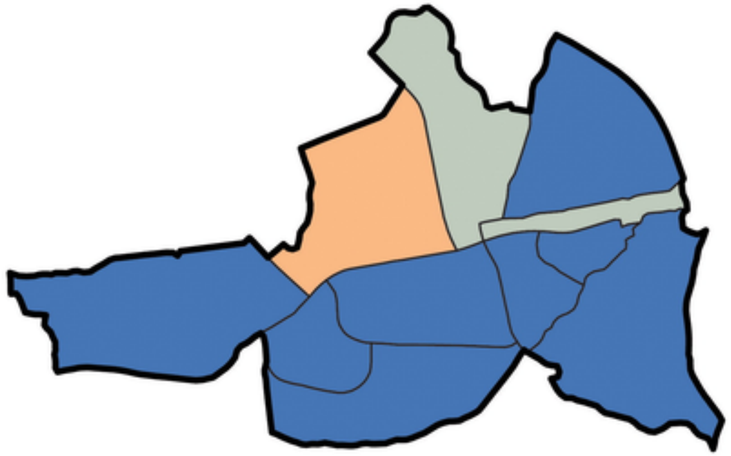
Openshaw, 1984

A common source of bias in spatial aggregation: MAUP

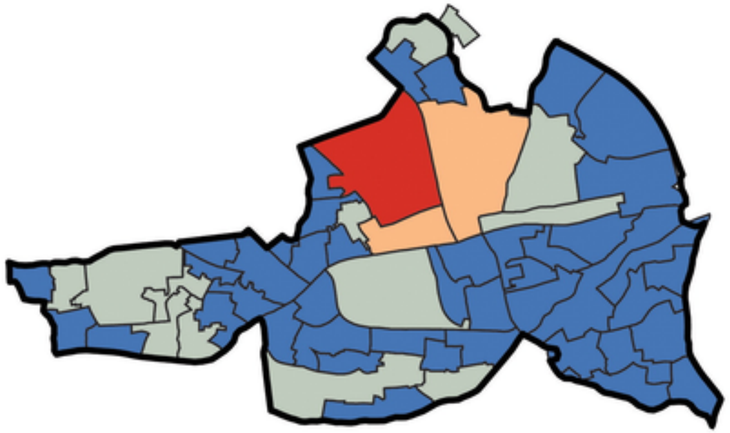


A common source of bias in spatial aggregation: MAUP

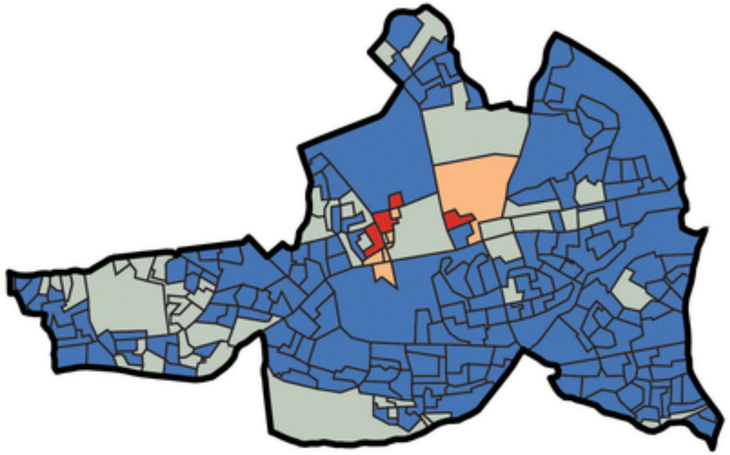
1. Electoral Divisions (EDs), 11 in Study Area



2. Enumerator Area (EAs), 56 in Study Area



3. Small Areas (SAs), 237 in Study Area

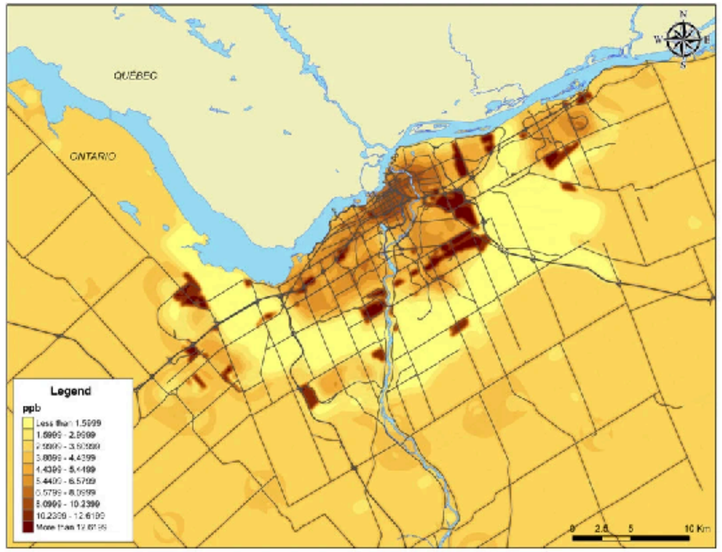


Modifiable areal unit problem (MAUP)
Study Area: Tallaght, Dublin
Housing Vacancy Rate, 2011

- < 5%
- 5% to < 15%
- 15% to < 30%
- 30% plus

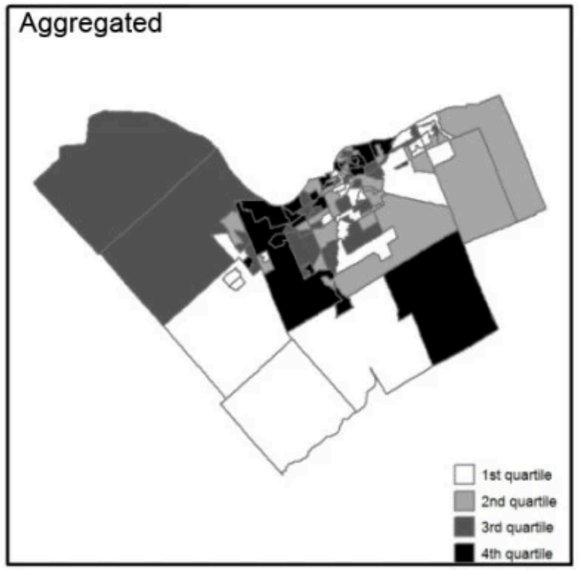
https://www.tandfonline.com/doi/full/10.1080/21681376.2014.983149

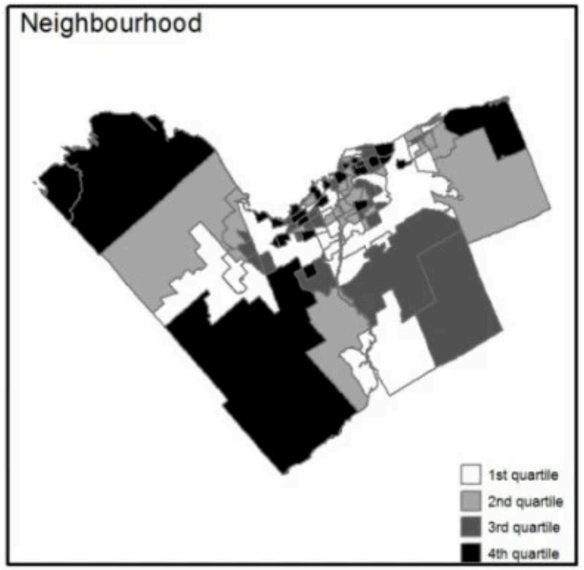
A common source of bias in spatial aggregation: MAUP

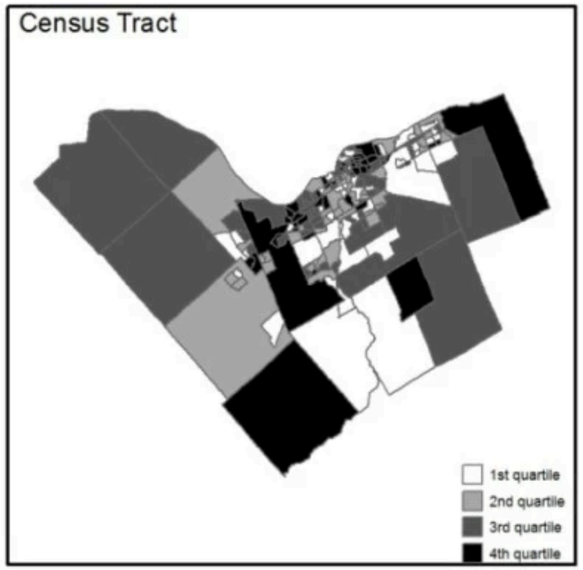


Pollution

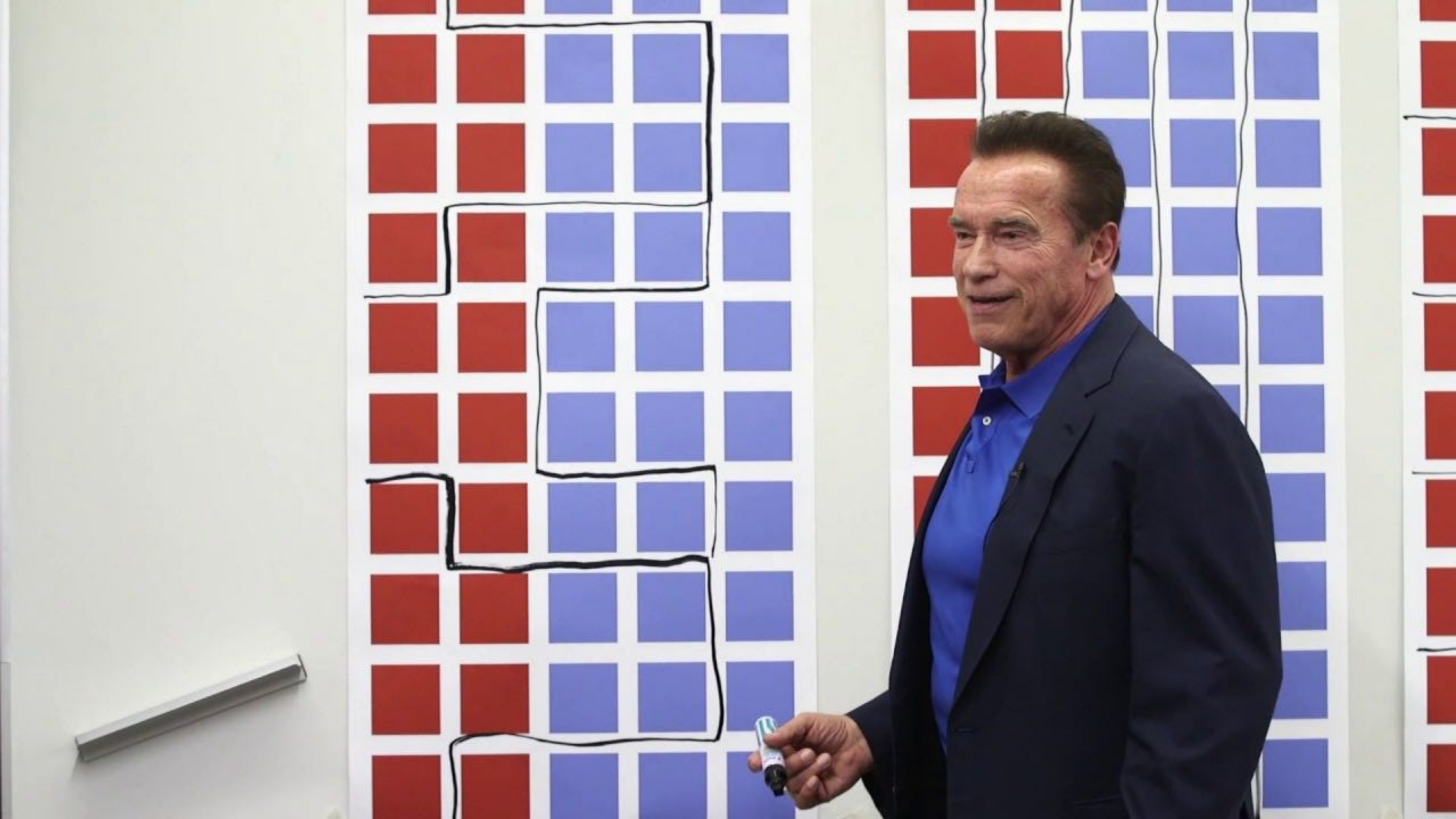
Morbidity



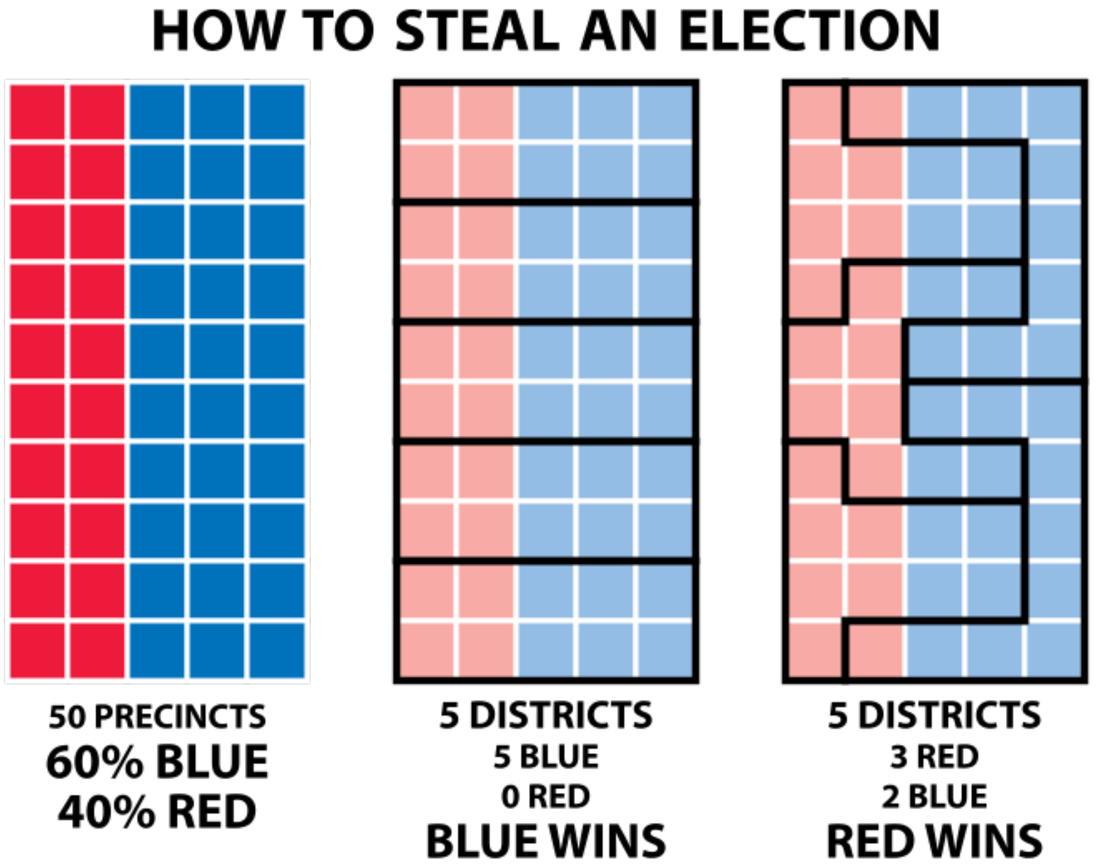


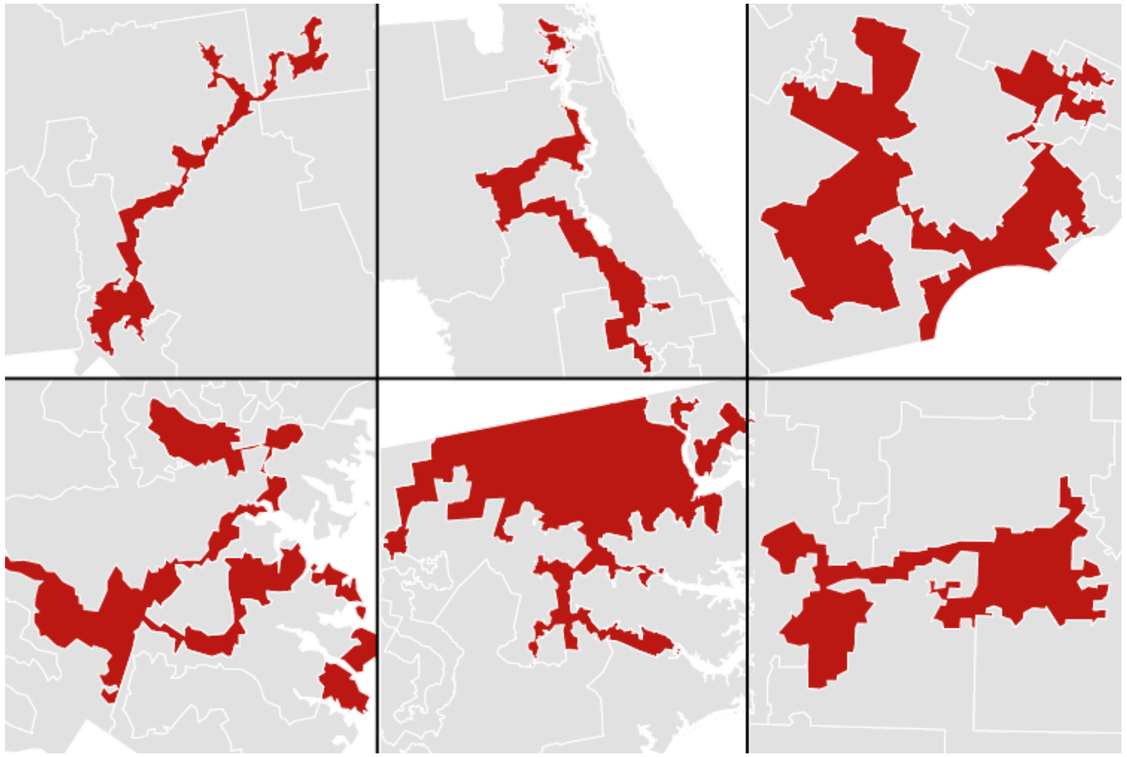


https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-10-58



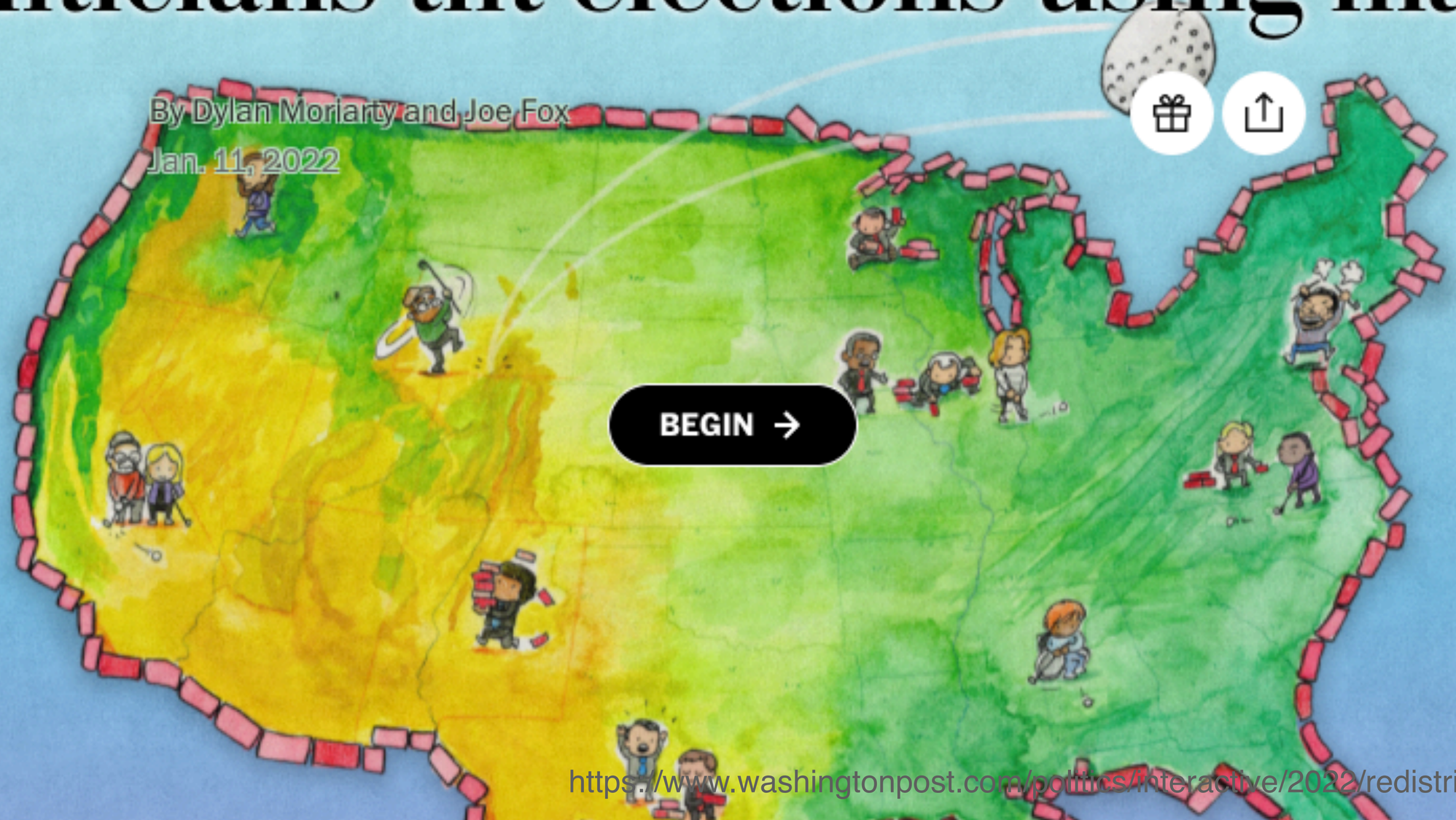
The MAUP is abused for Gerrymandering





Play mini golf to see how politicians tilt elections using maps

By Dylan Moriarty and Joe Fox
Jan. 11, 2022



https://www.washingtonpost.com/politics/interactive/2022/redistricting-mini-golf

Can You Gerrymander Your Party to Power?

By Ella Koeze, Denise Lu and Charlie Smart

Gerrymandering is the intentional distortion of political districts to give one party an advantage, and it has been criticized for disenfranchising many voters and fueling deeper polarization.

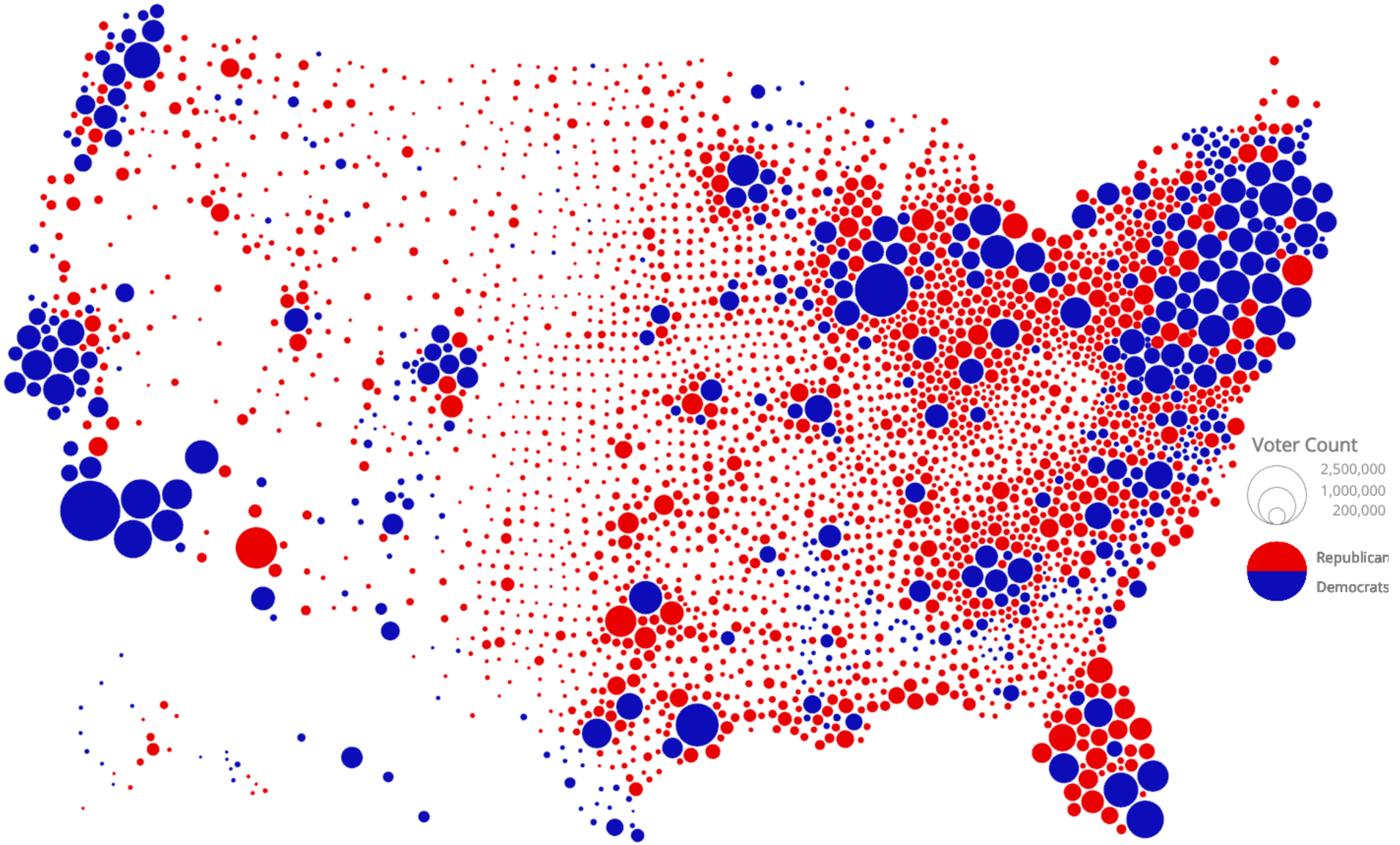
To help you understand it better, we created an imaginary state called **Hexapolis**, where your only mission is to gerrymander your party to power.

https://www.nytimes.com/interactive/2022/01/27/us/politics/congressional-gerrymandering-redistricting-game-2022.html?unlocked_article_code=AAAAAAAAAAAAAAAACFIPuonLyYiZ_tU1Gw5CBWySB4B99Fre3L-

Related bias: Area is not population

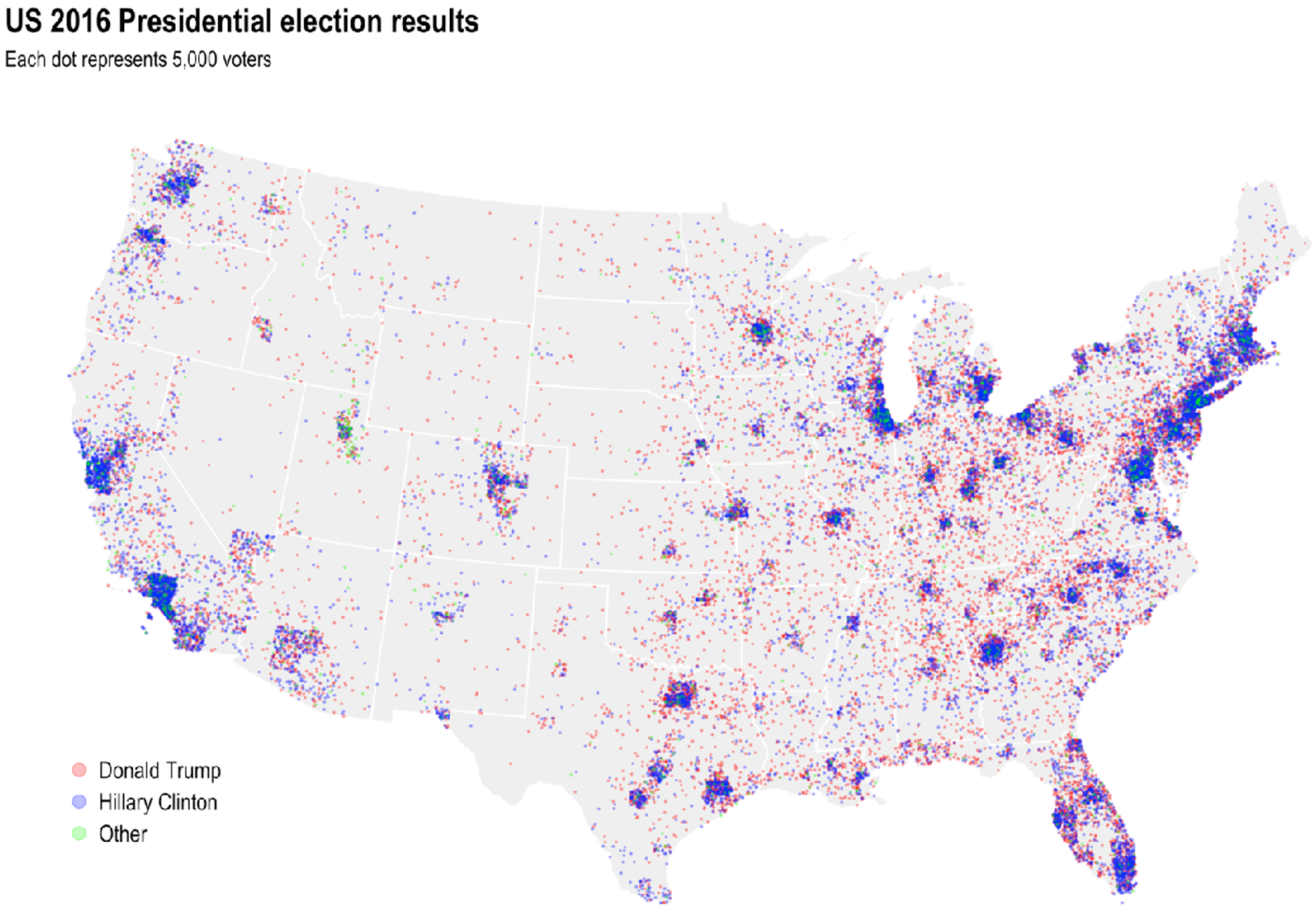
http://try-to-impeach-this.jetpack.ai/

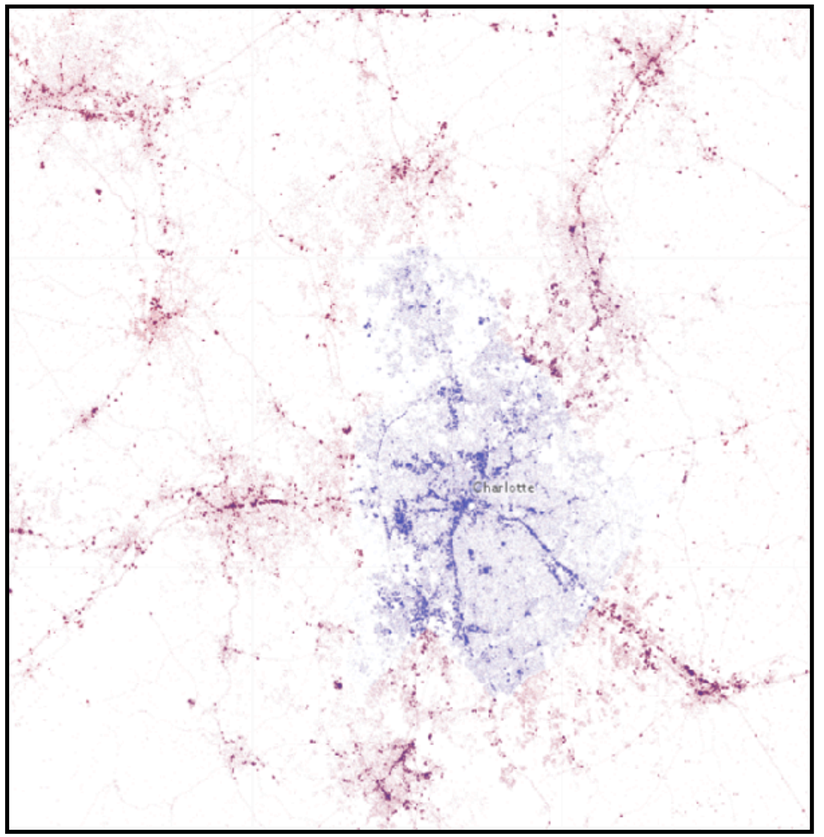
Related bias: Area is not population



http://try-to-impeach-this.jetpack.ai/

A dot density map can reduce the bias





https://www.andybeger.com/2018/05/11/dot-density-map-of-the-2016-election/
https://www.maproomblog.com/2018/04/kenneth-fields-dot-density-election-map-redux/

Always mind the MAUP when exploring aggregated data

Classification = Transforming continuous into categorical

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

reject
accept
accept

How many categories should there be? If 2: Binarization
How should the values be mapped? What should be the thresholds?

Number of finds in an excavation grid

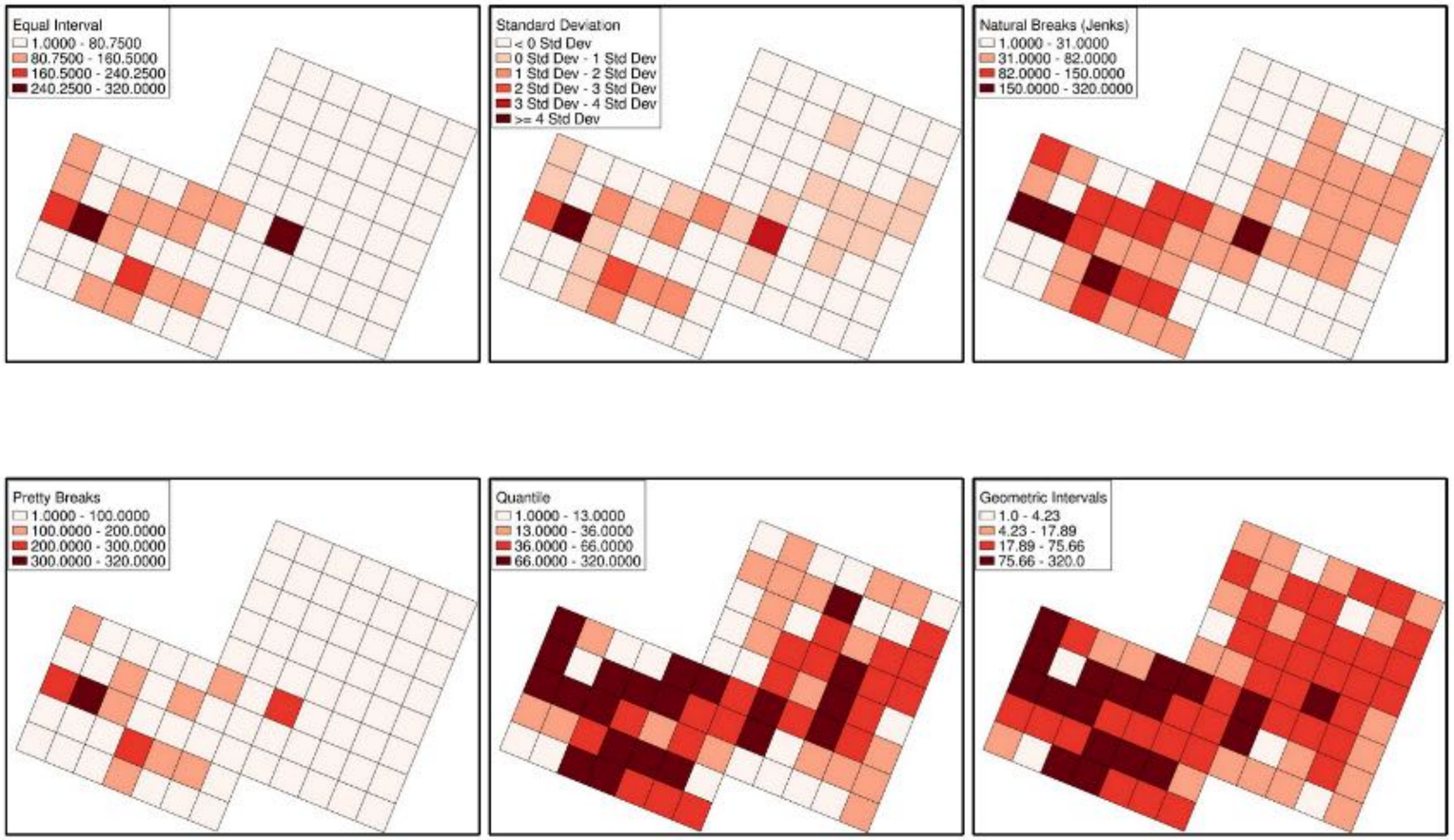
Group experiment

Each of 3 groups describes their picture in 1 sentence.

Focus: West vs East

Data can be classified very differently

Example: Weight of finds in an excavation grid



Same data, different split points

http://184.106.205.13/uselessarchaeology/CMSimple_UA/?Programming___QGIS_Geometric_Class

Choropleth maps visualize geospatial data with colors

Choro = region
pleth = multiple

Combine: Polygons + Attribute values

Values are often grouped into classes to reduce complexity

Choropleth maps visualize geospatial data with colors

Choro = region
pleth = multiple

Combine: Polygons + Attribute values

Values are often grouped into classes to reduce complexity

How you do this depends on your:
- Audience
- Goal (highlight outliers, show distribution, deceive,..)

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

Formally, you must select thresholds to classify your data

The classification problem is to define class boundaries such that

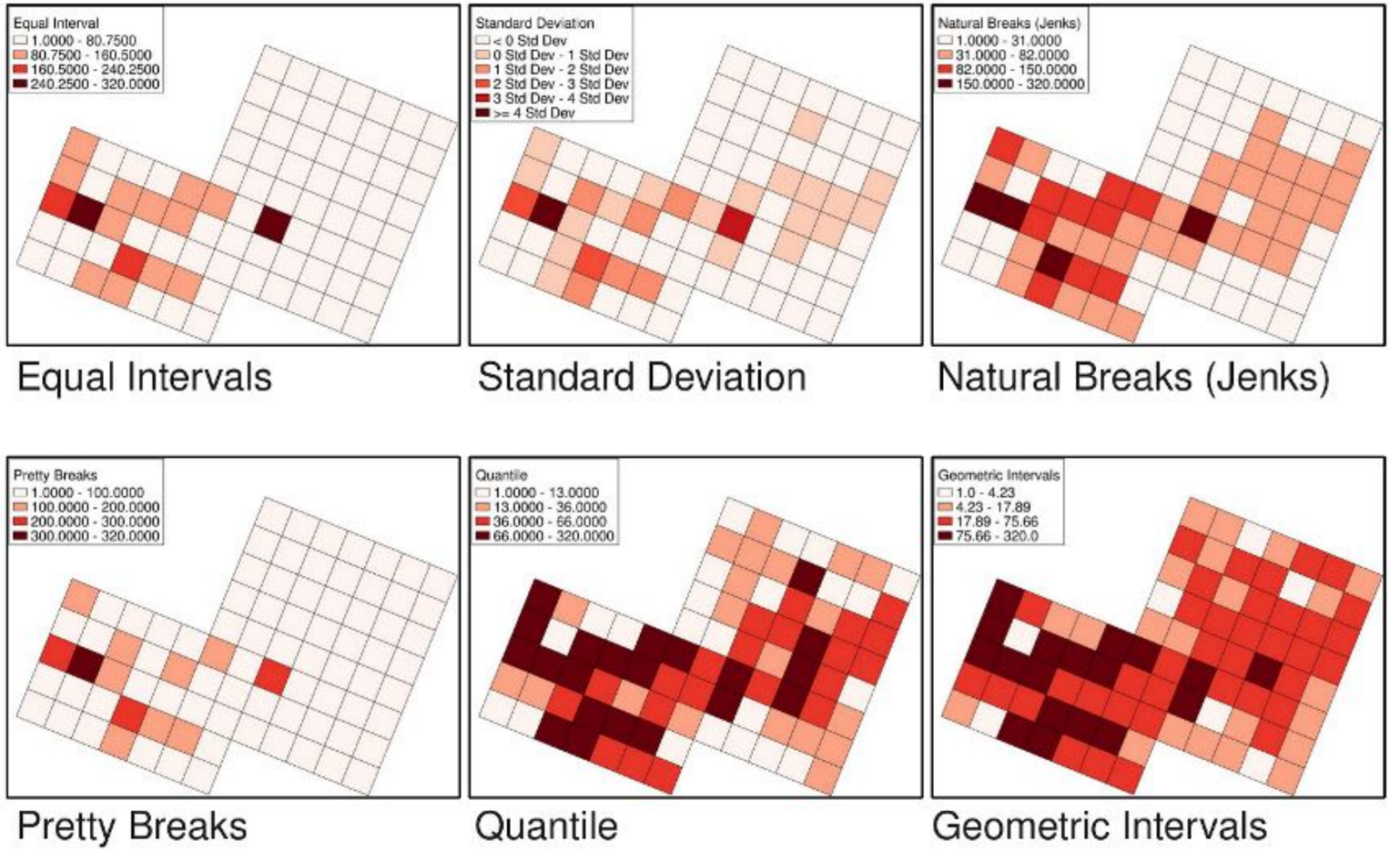
$$c_j < y_i \leq c_{j+1} \; \forall y_i \in C_j$$

where y_i is the value of the attribute for spatial location i,
j is a class index,
c_j represents the lower bound of interval j.

The choice of class boundaries defines the classification scheme

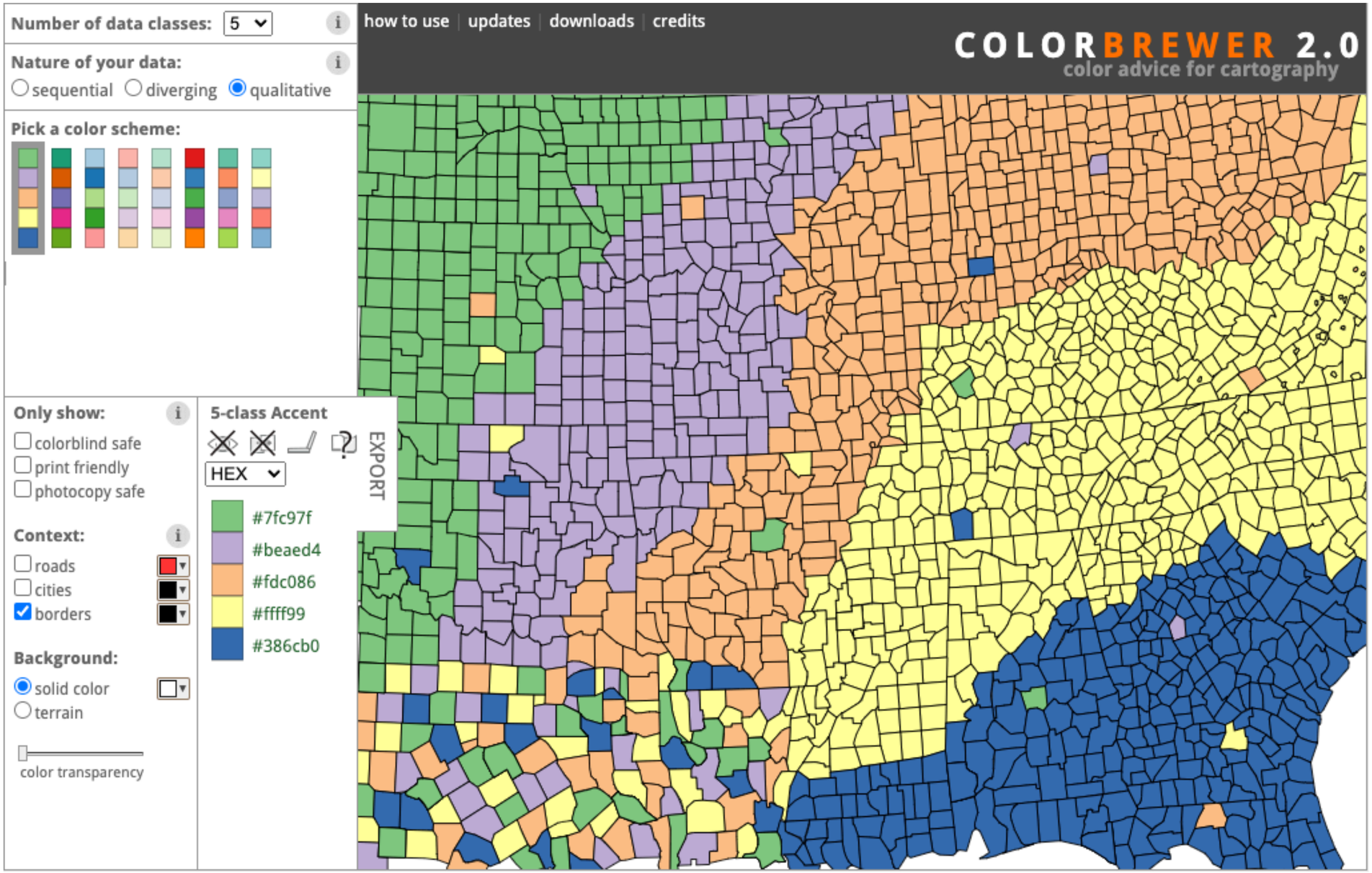
mapclassify:

'Head-Tail Breaks',
'Fisher-Jenks',
'Maximum Breaks',
'Equal Interval',
'MaxP',
'Quantiles',
'Jenks Caspall',
'Mean-Standard Deviation'



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

For choice of colors use tools like color brewer

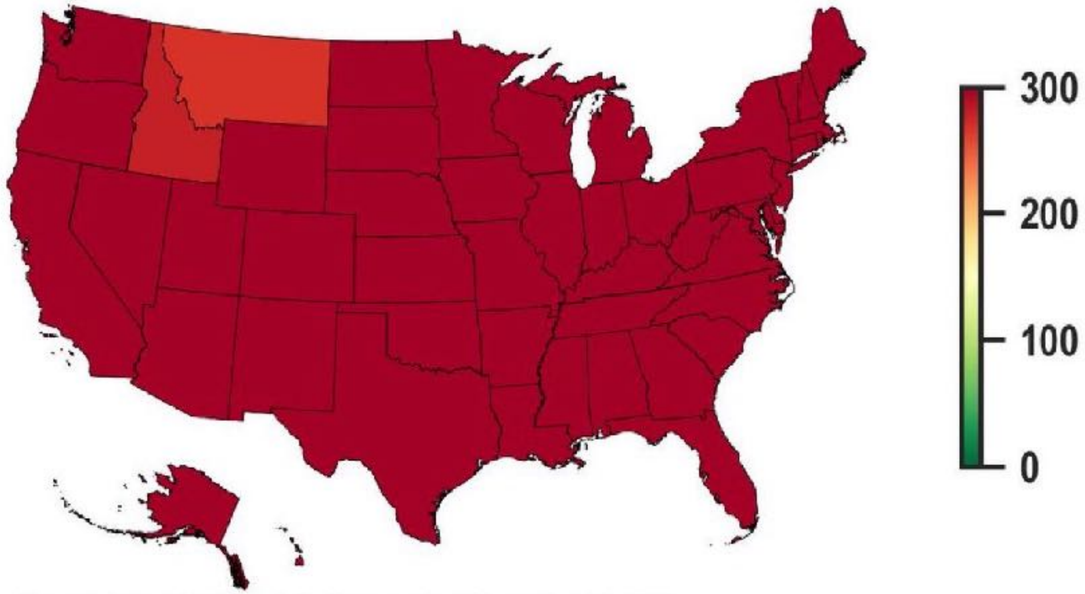


https://colorbrewer2.org/

Example: Wrong use of colors and scales

How the year started

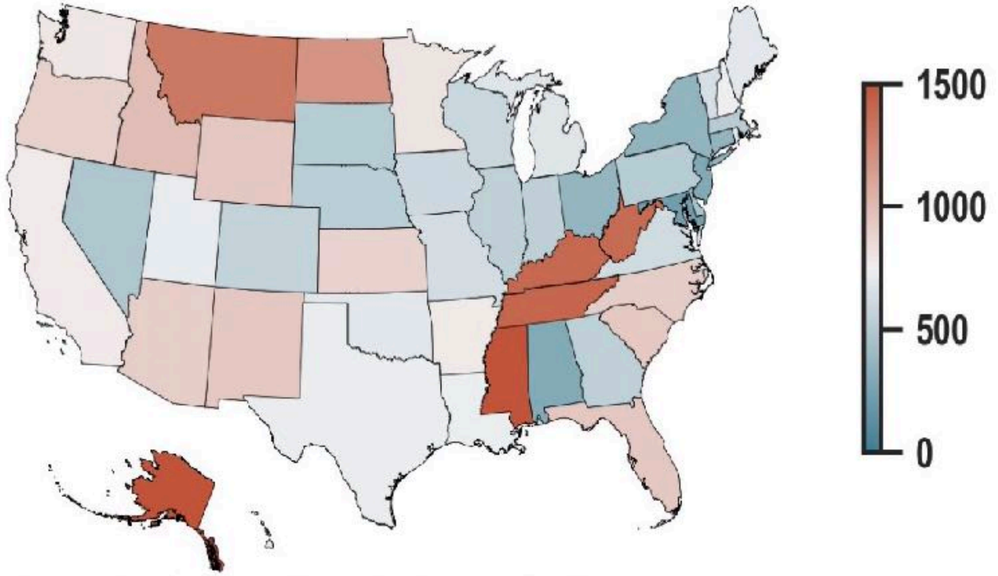
New COVID-19 cases per million population



Source: Johns Hopkins, J. P. Morgan. As of December 31, 2021

How it is going

New COVID-19 cases per million population



Source: Johns Hopkins, J. P. Morgan. As of February 08, 2022

https://twitter.com/carlquintanilla/status/1491824980648247296?s=20

Jupyter

Sources and further materials for today's class

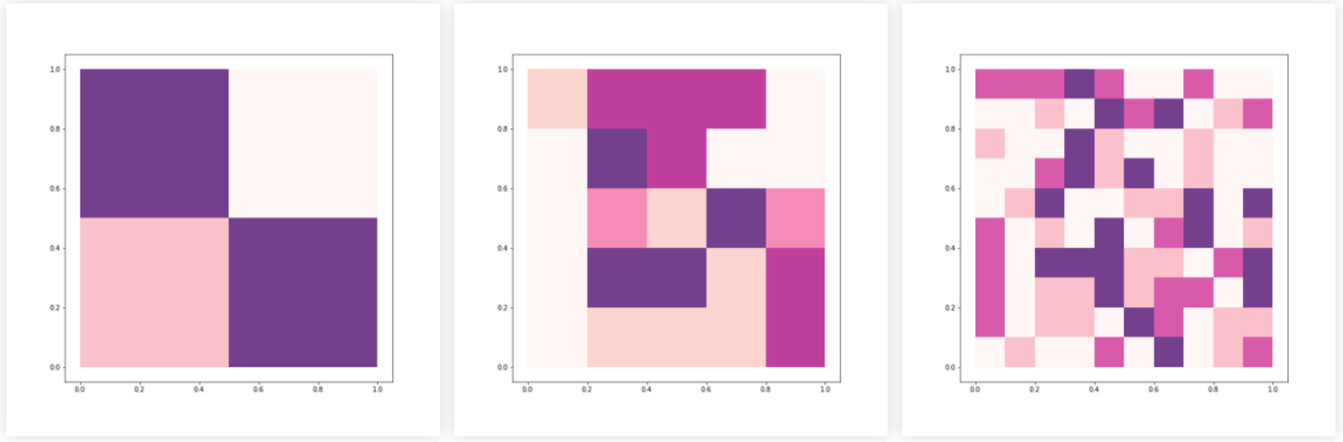
Geographic Data Science with Python

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

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

Take home messages for today

The choice of scale (MAUP) and classification scheme allows to create *arbitrary* interpretations



Aggregated data is simpler,
but can be very biased