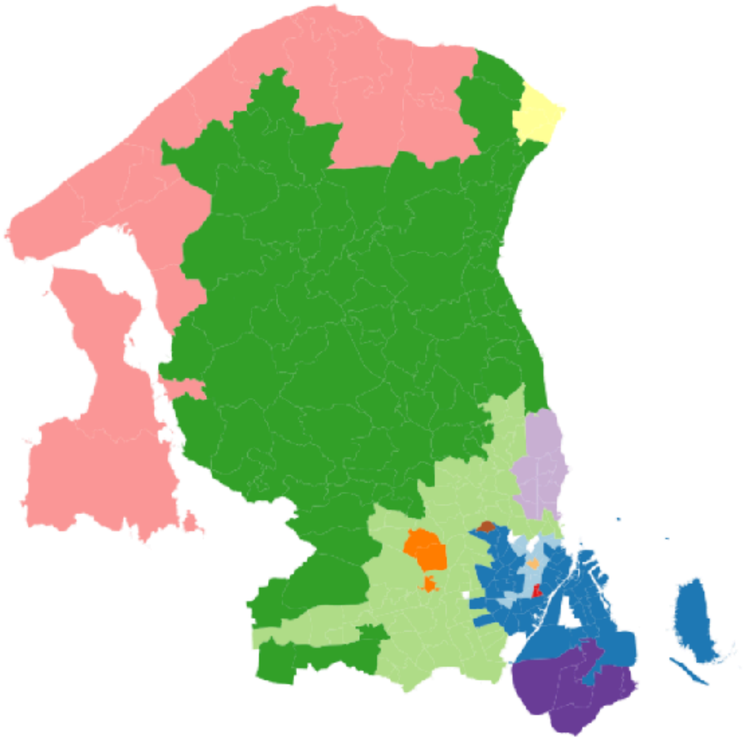
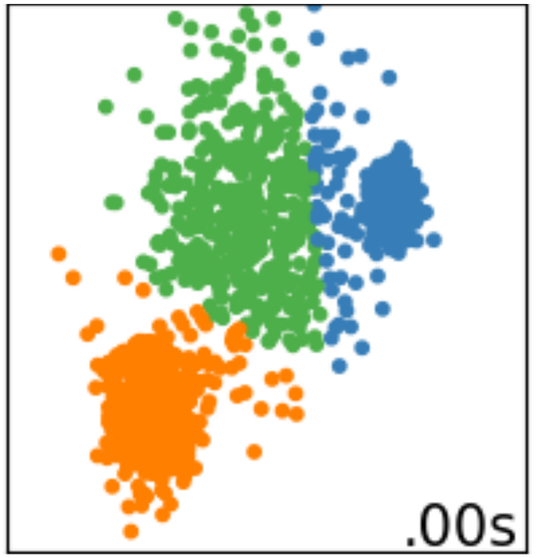
Lecture 6: Spatial clustering

Instructor: Ane Rahbek Vierø
Mar 6, 2023

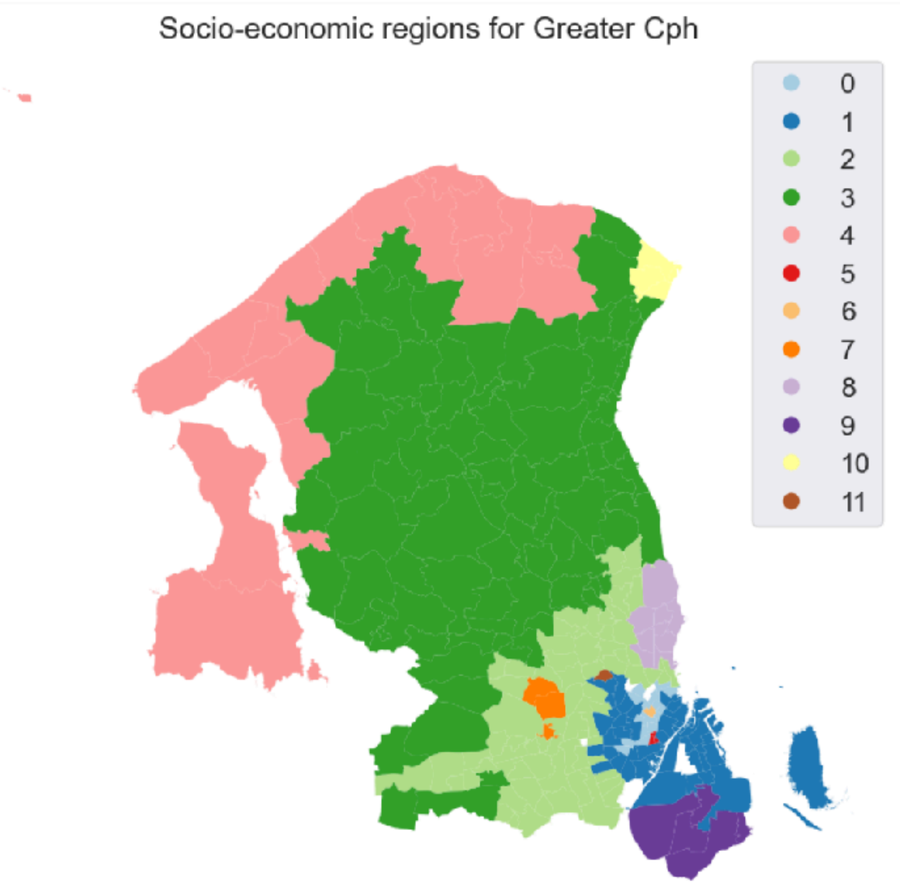


Today you will learn about spatial clustering

Clustering

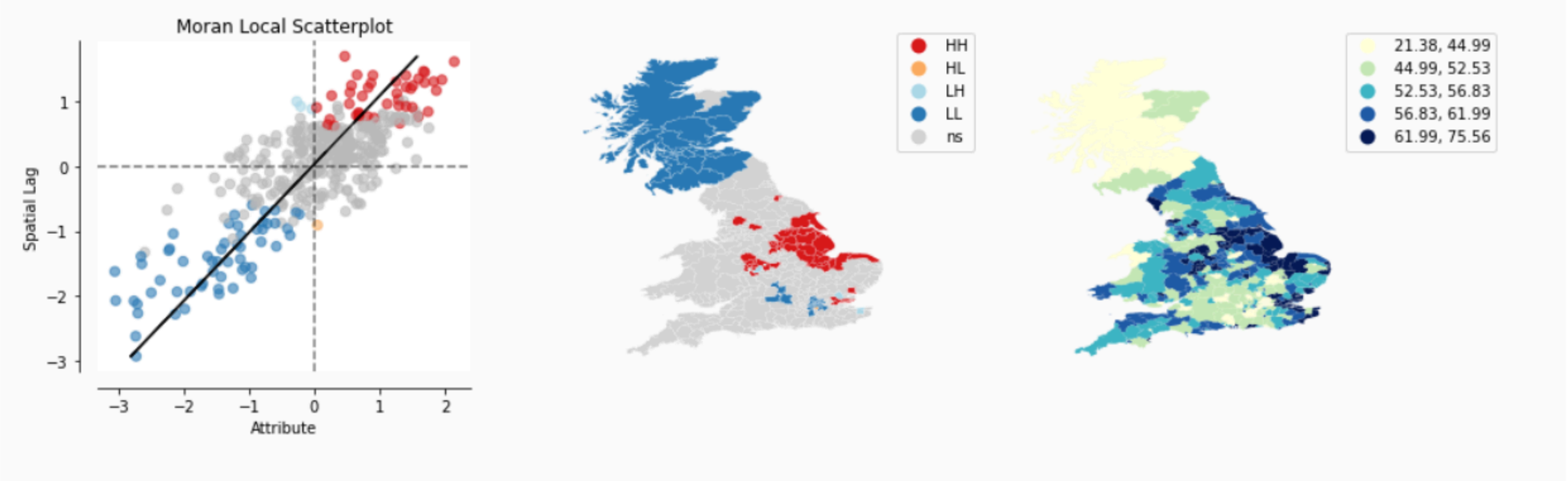


Regionalization with socioeconomic data



$$IPQ_i = \frac{A_i}{A_c} = \frac{4\pi A_i}{P_i^2}$$

Last week: Spatial Autocorrelation



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

Everything should be made as simple as possible, but not simpler.

Albert Einstein

From univariate to multivariate analysis

The world is complex and multidimensional.

Univariate	Multivariate
Percent of foreign-born	Neighborhood
Years of schooling	Human development
Monthly income	Deprivation

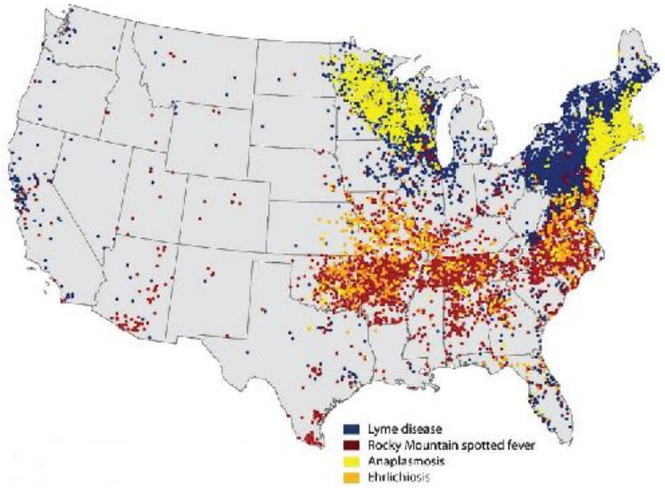
Cluster analysis

Cluster analysis is the division of data into groups that are meaningful, useful, or both.

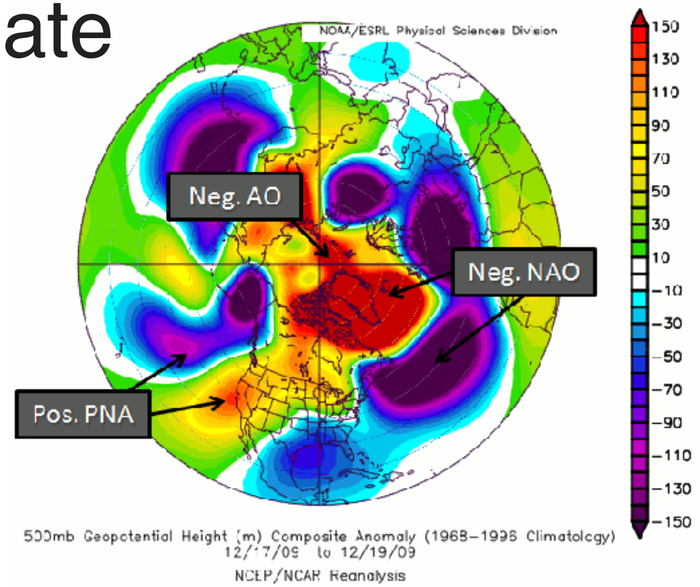
It simplifies, and can be the starting point for other purposes like data summarization.

Cluster analysis: useful for understanding/handling data

Medicine



Climate



Business

Segment customers for additional analysis and to target for marketing activities

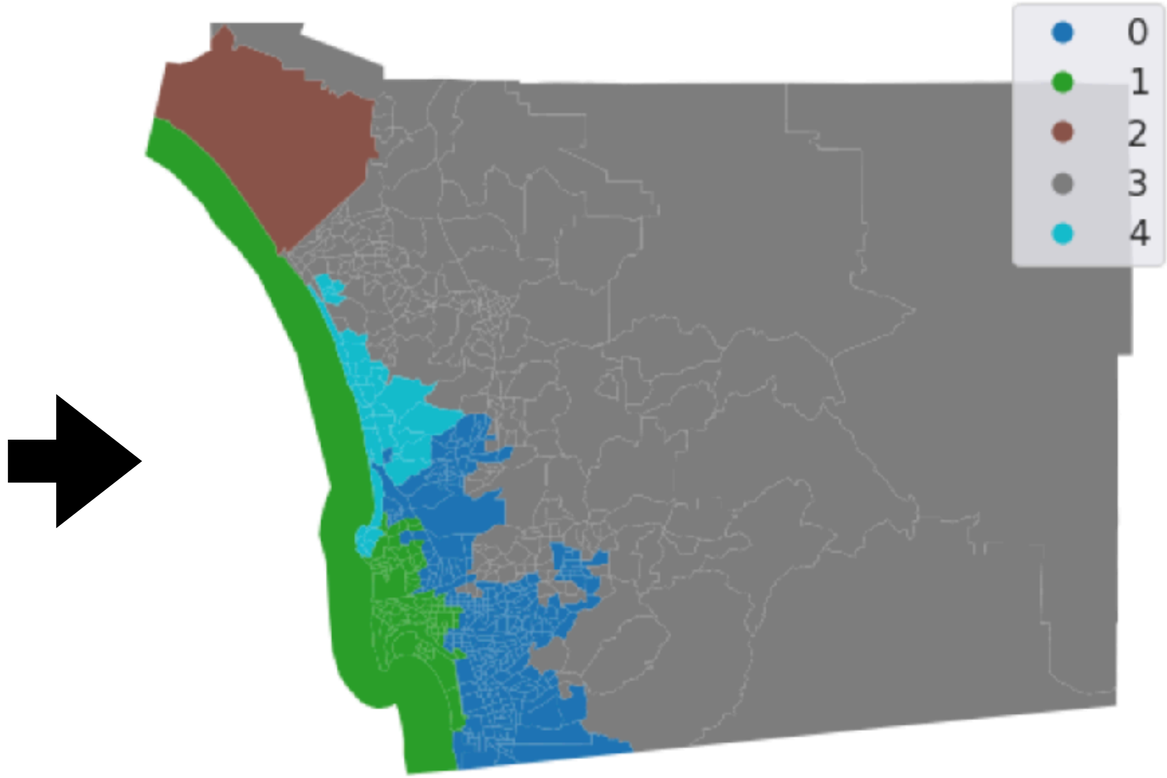
Facebook

Facebook allowed advertisers to target 'Jew haters'

Embarrassing discovery that Facebook let advertisers target users interested in antisemitic topics comes as the social network's ad practices are under scrutiny

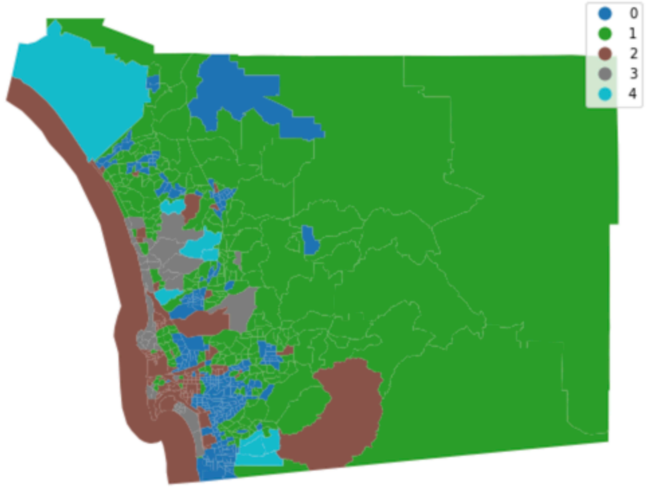
Clustering helps us make sense of complex spatial patterns

Clustering helps us make sense of complex spatial patterns

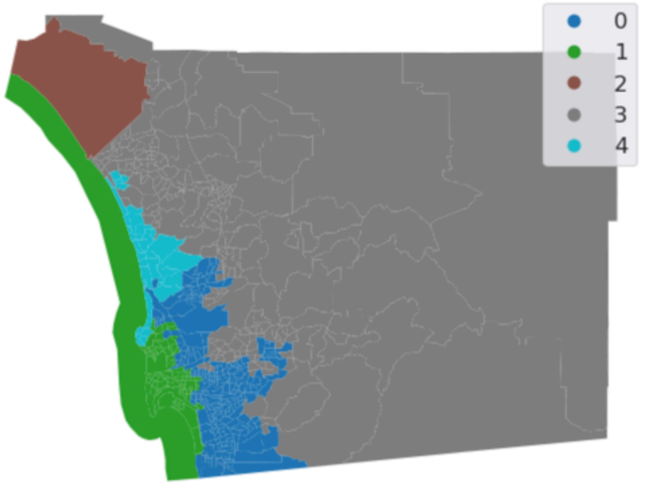


Types of clustering

Non-spatial

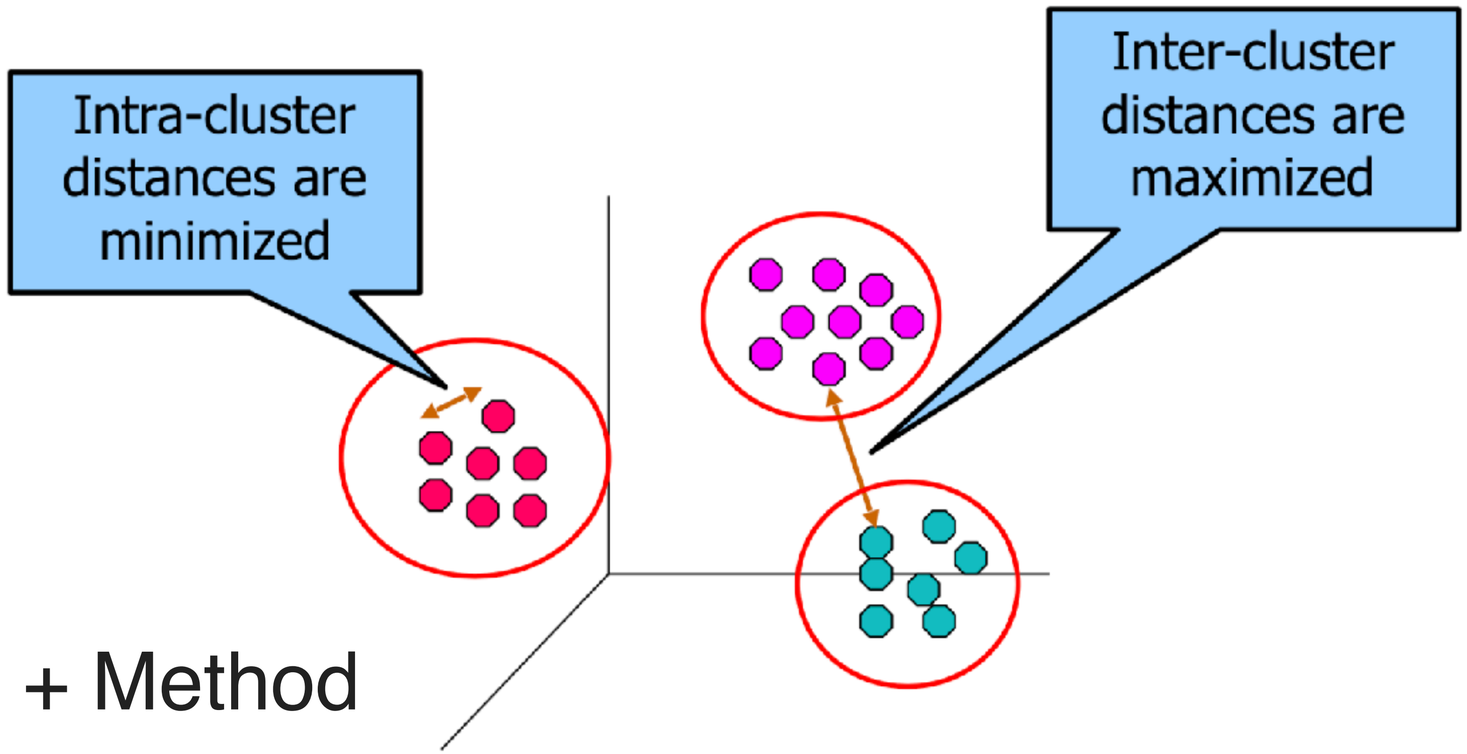


Spatial / Regionalization



Cluster analysis: General concept

Finding groups of objects such that the objects in a group will be similar or related to one another and dissimilar or unrelated to the objects in other groups



Clustering = Distance + Method

Clustering requires scaled data

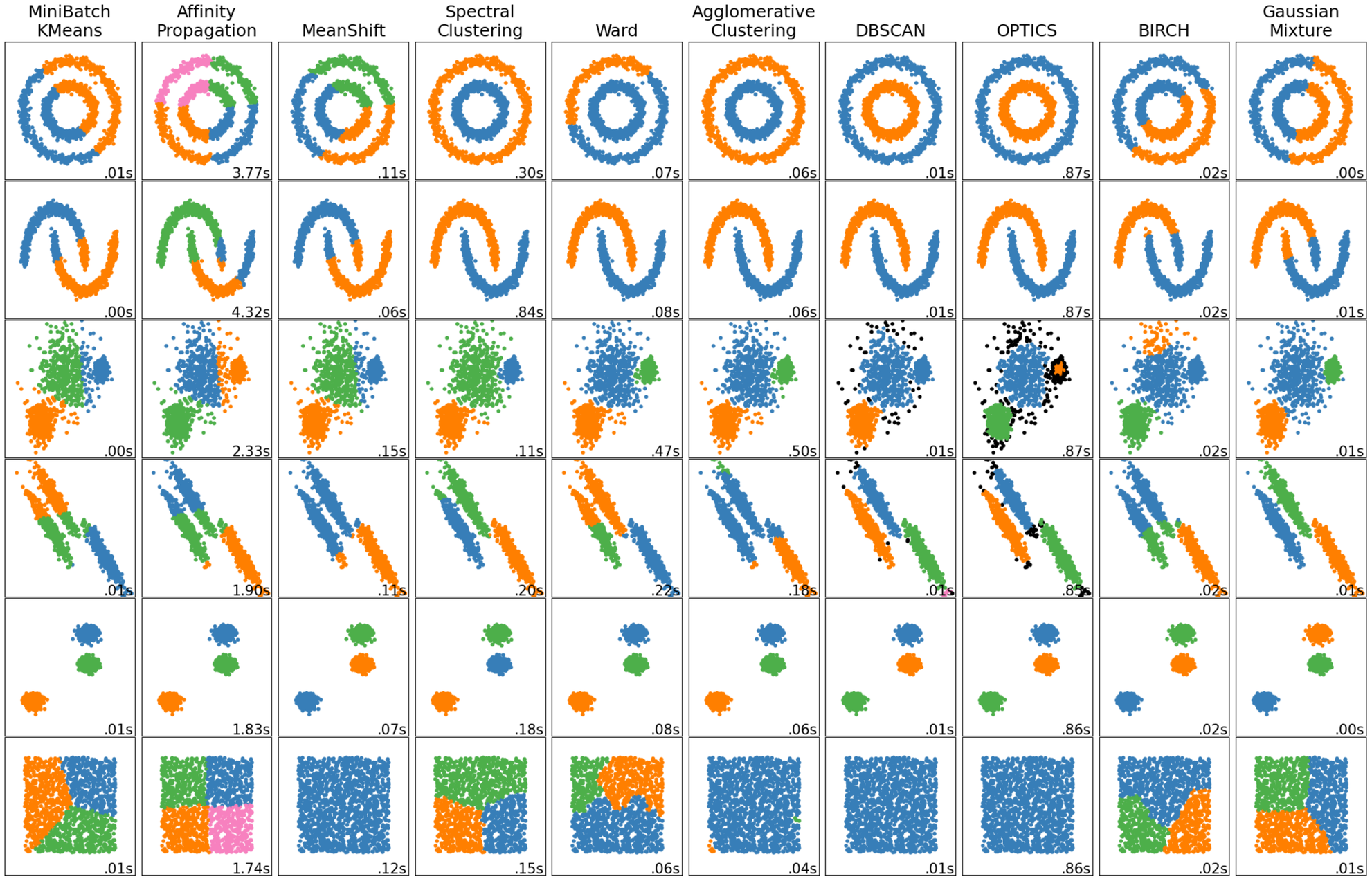
Clustering requires scaled data

Scale (standardize)

$$z = \frac{x_i - \bar{x}}{\sigma_x}$$

Robust Scale

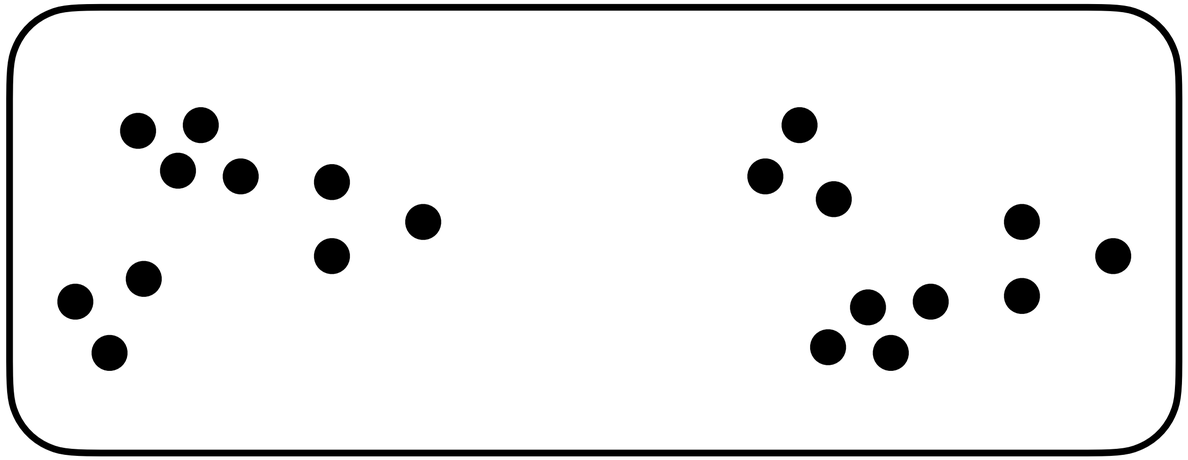
$$z = \frac{x_i - \tilde{x}}{\lceil x \rceil_{75} - \lceil x \rceil_{25}}$$

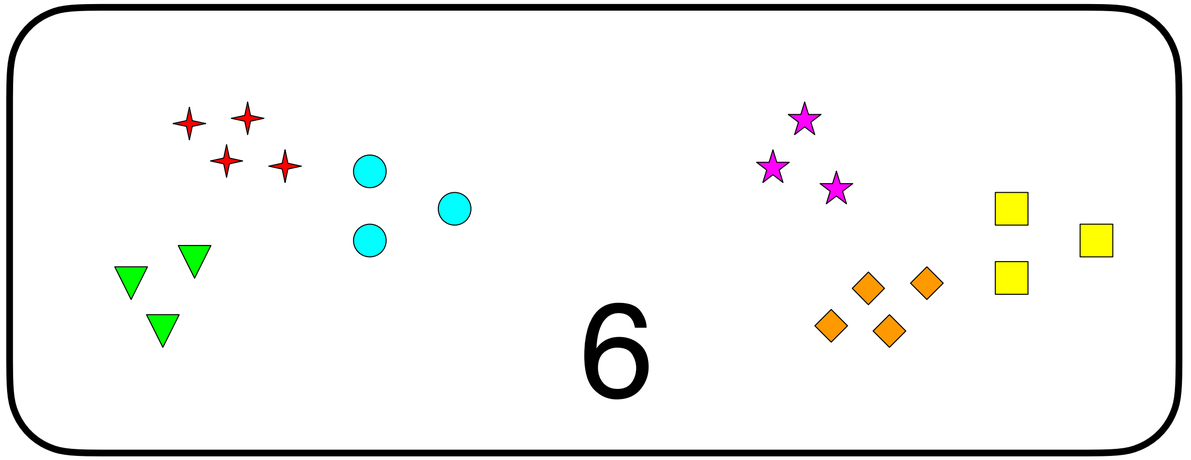


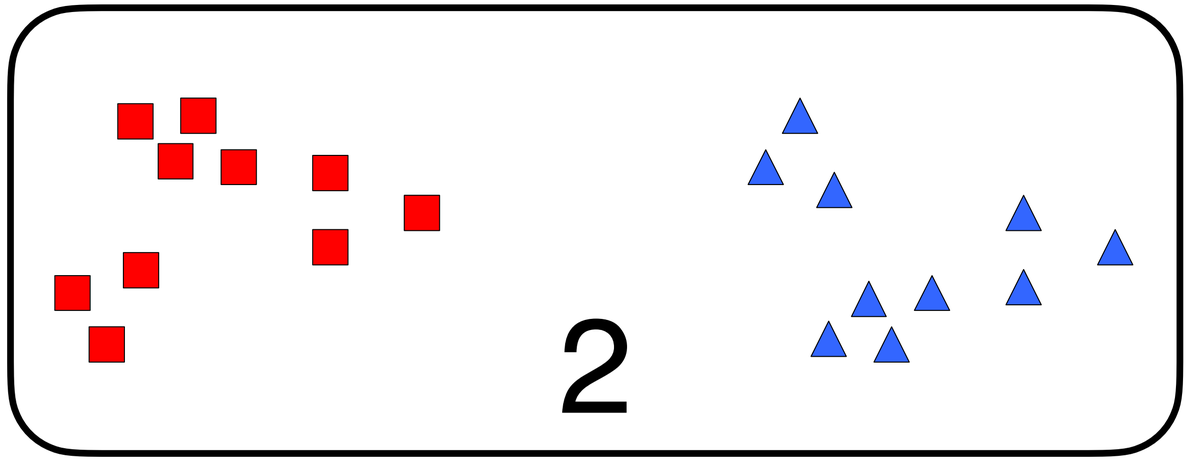
https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods

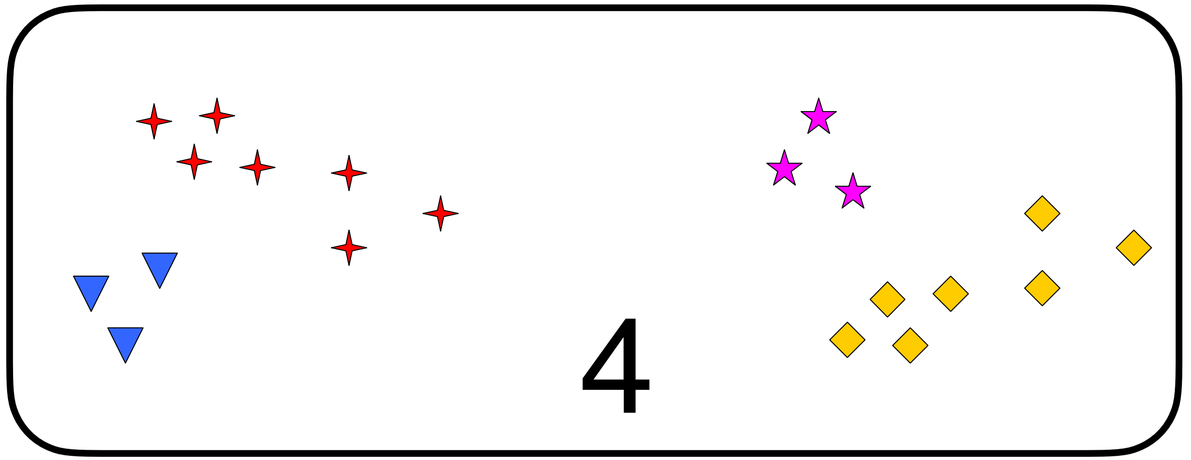
How many clusters are there?

Clustering is ambiguous







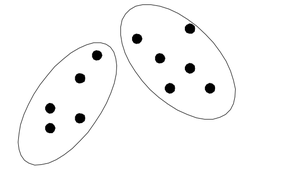


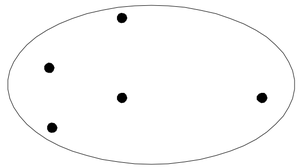
There is no clustering technique that is universally applicable in uncovering the variety of structures present in multidimensional data sets

Partitional vs hierarchical

Partitional clustering

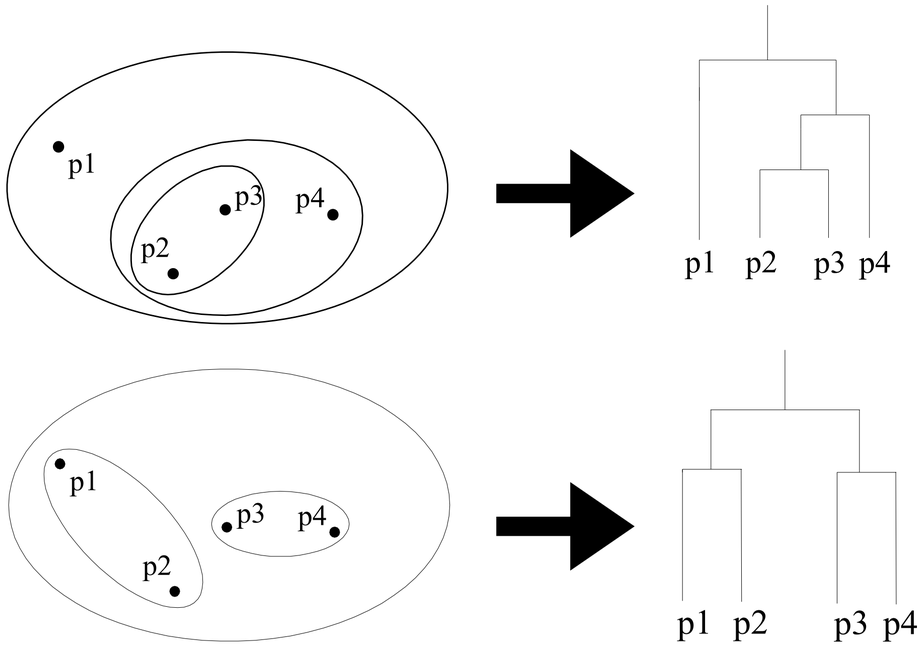
A division into non-overlapping subsets. Each data object is in exactly one subset.





Hierarchical clustering

A set of nested clusters organized as a tree



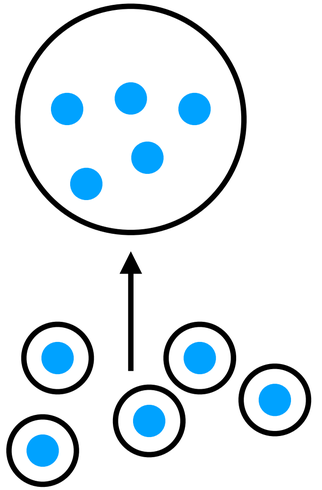
Hierarchical clustering

Agglomerative clustering (bottom-up)

Start with points being individual clusters

At each step merge the closest pair of clusters.

Needs: Similarity measure

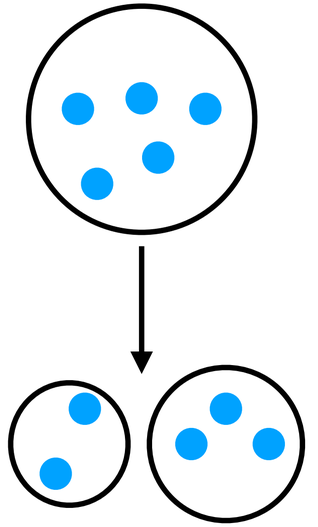


Divisive clustering (top-down)

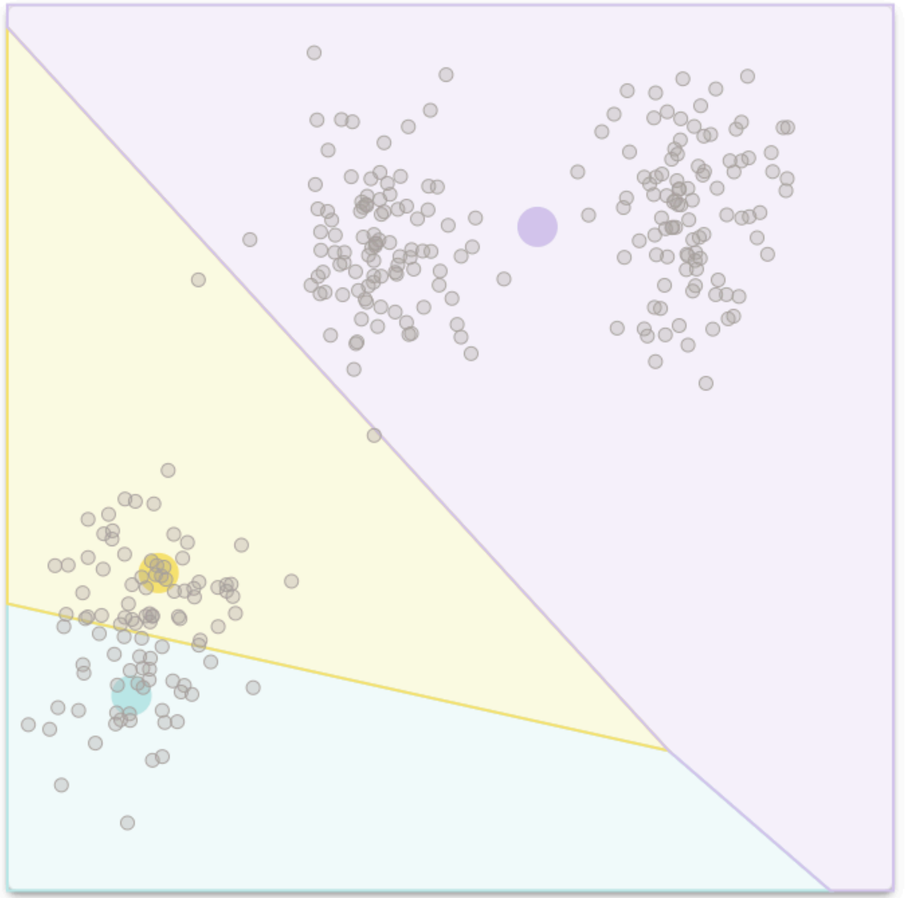
Start with one all-inclusive cluster.

At each step split a cluster until clusters of single points remain.

Needs: Way to decide a split



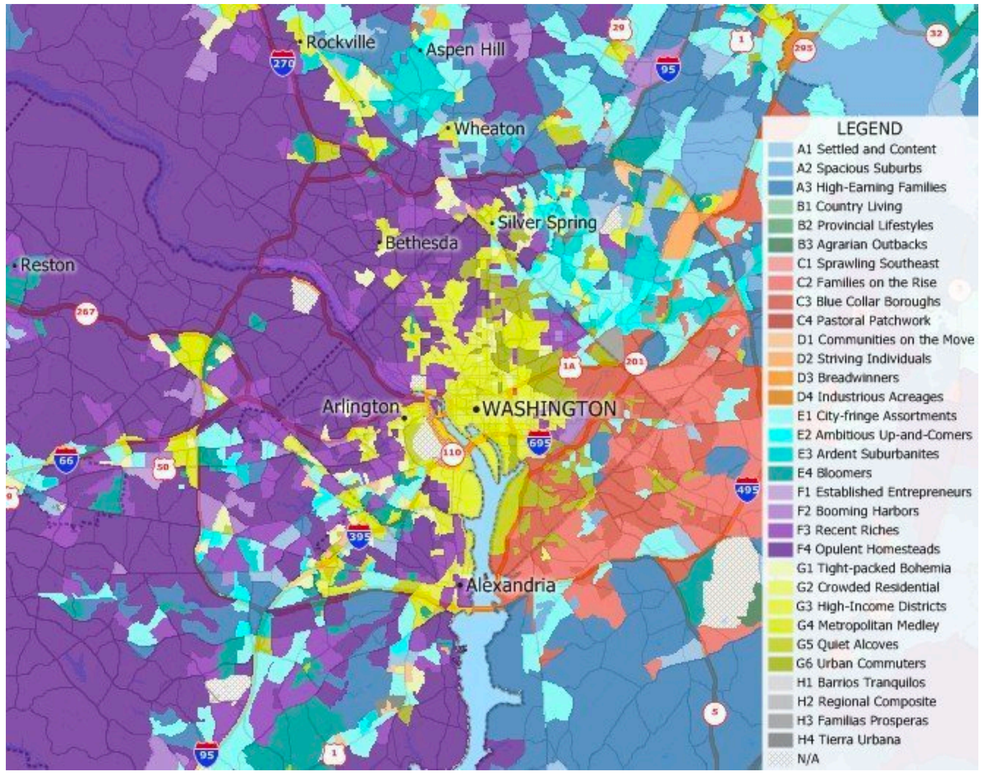
K-means



https://k-means-explorable.vercel.app/

Geodemographic analysis

Geodemographic analysis aims to identify similar “neighborhoods" based on people and place characteristics



Does our data set even have clusters?

A clustering algorithm will do its job and give us clusters.

But does it make sense?

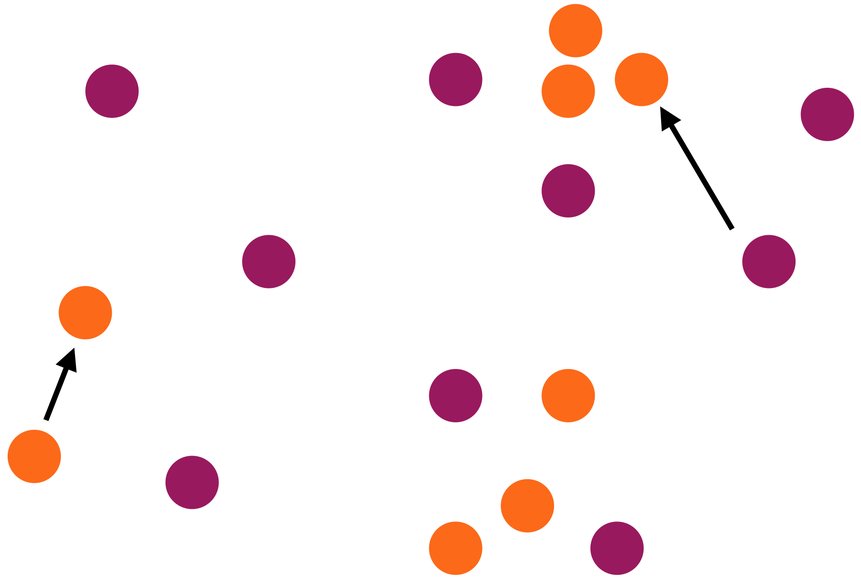
The Hopkins statistic H measures clustering tendency

A bit like Moran's I, but usually used in a non-spatial context

The Hopkins statistic H measures clustering tendency

1) Generate p points randomly distributed across the data space, and 2) sample p data points.

For both sets, find the distance to the nearest neighbors in the original data set.



The Hopkins statistic H measures clustering tendency

Generate p points randomly distributed across the data space, and sample p data points.

For both sets, find the distance to the nearest neighbors in the original data set.

Hopkins statistic $$H = \frac{\sum_{i=1}^{p} u_i}{\sum_{i=1}^{p} w_i + \sum_{i=1}^{p} u_i}$$

distance of synthetic data

distance of real data

H=1: real data is highly clustered

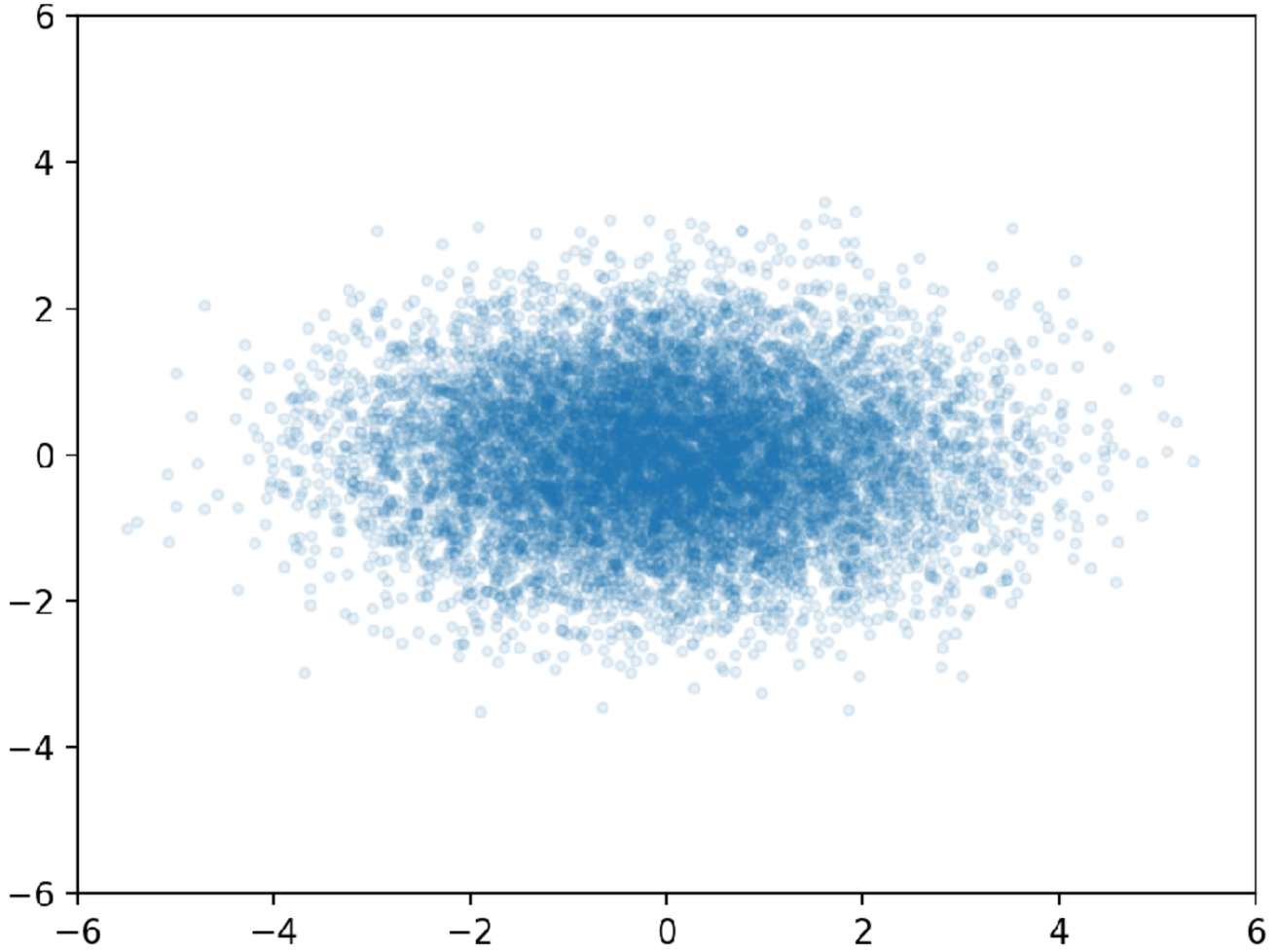
H>0.75: real data is clustered at 90% confidence level

H=0.5: random and real points have same distance distributions

H≅0: real data is uniformly distributed

The Hopkins statistic compares to a uniform distribution

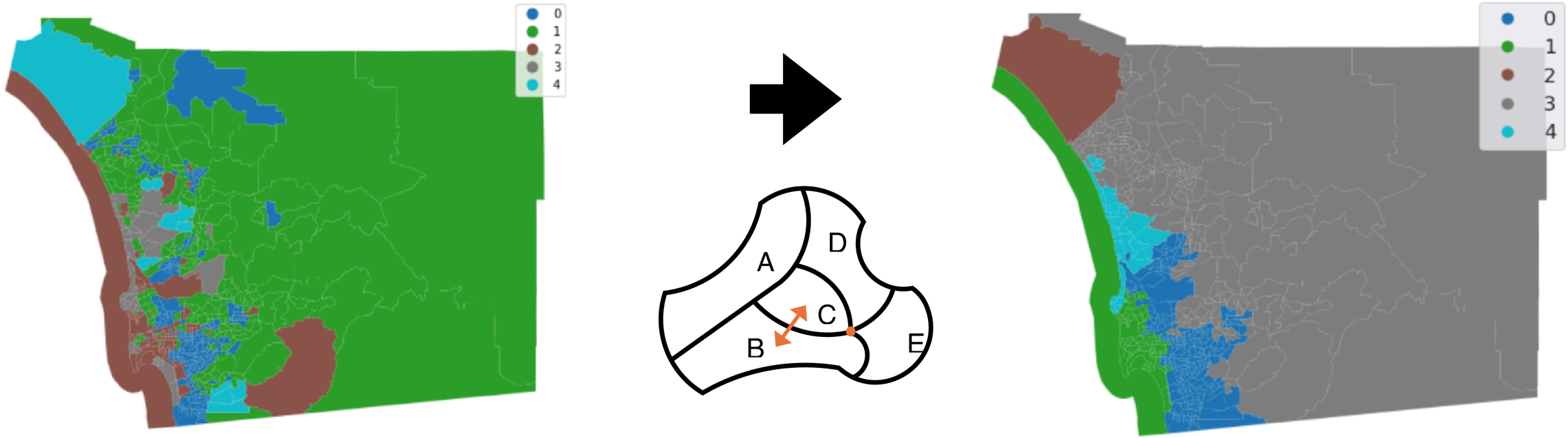
Criticism of H:
If there is just one "cluster", H is also close to 1



Spatially constrained clustering

Regionalization is "geographic clustering"

Regionalization is spatially constrained clustering where observations must be geographical neighbors to be able to be in the same cluster



Spatial contiguity constraint

You must be able to travel between all members in the cluster without traveling through non-member areas

Regionalization is "geographic clustering"

Regionalization works by aggregation

Predefined number of regions

Spatial contiguity constraint

Number of regions < number of areas

No overlapping regions

No empty regions

Regionalization is "geographic clustering"

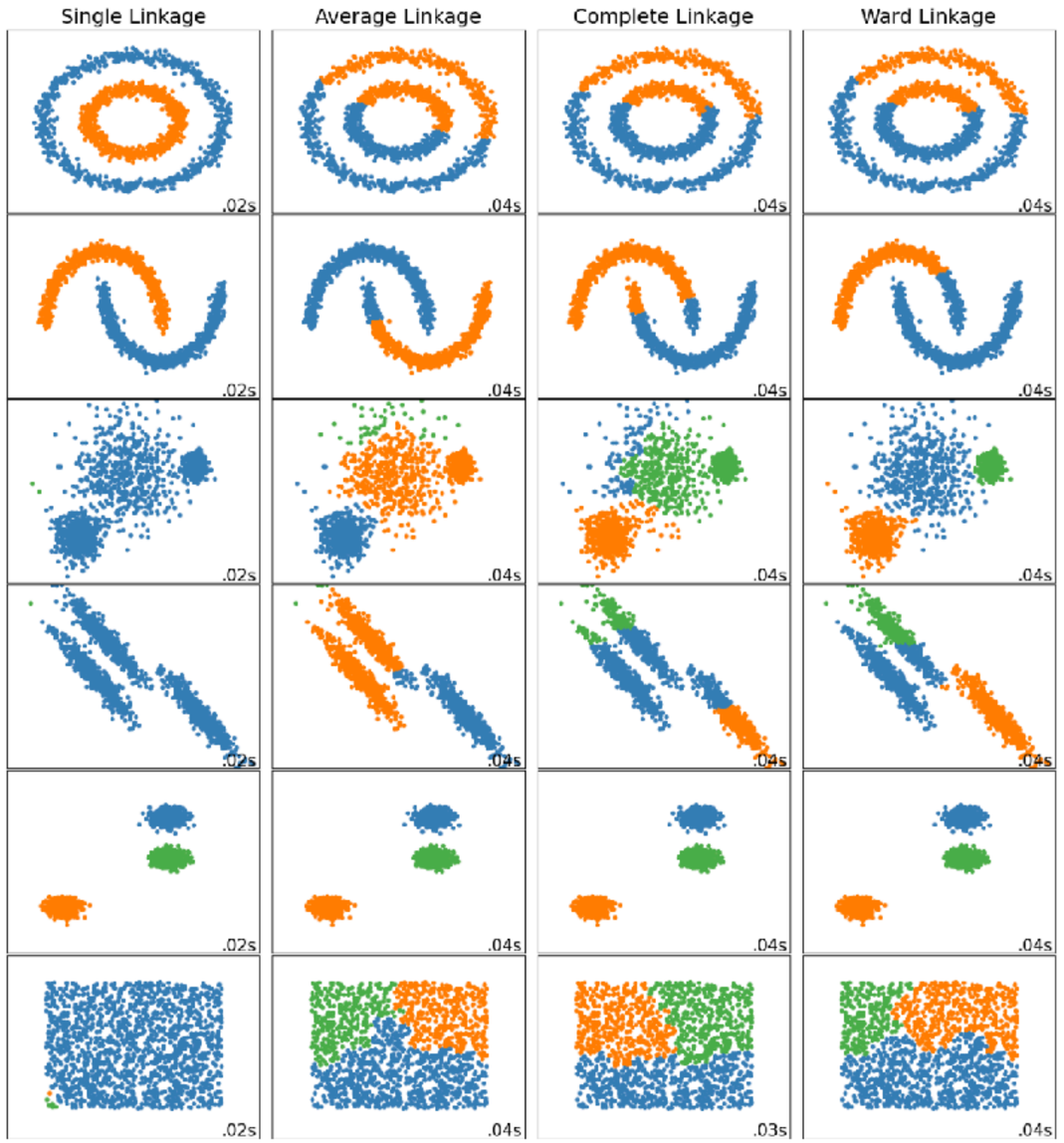
Regionalization results in analytical or functional regions

An alternative to regular aggregation

Less prone to MAUP, ecological fallacy, etc.

Agglomerative hierarchical clustering (AHC)

…with contiguity constraint



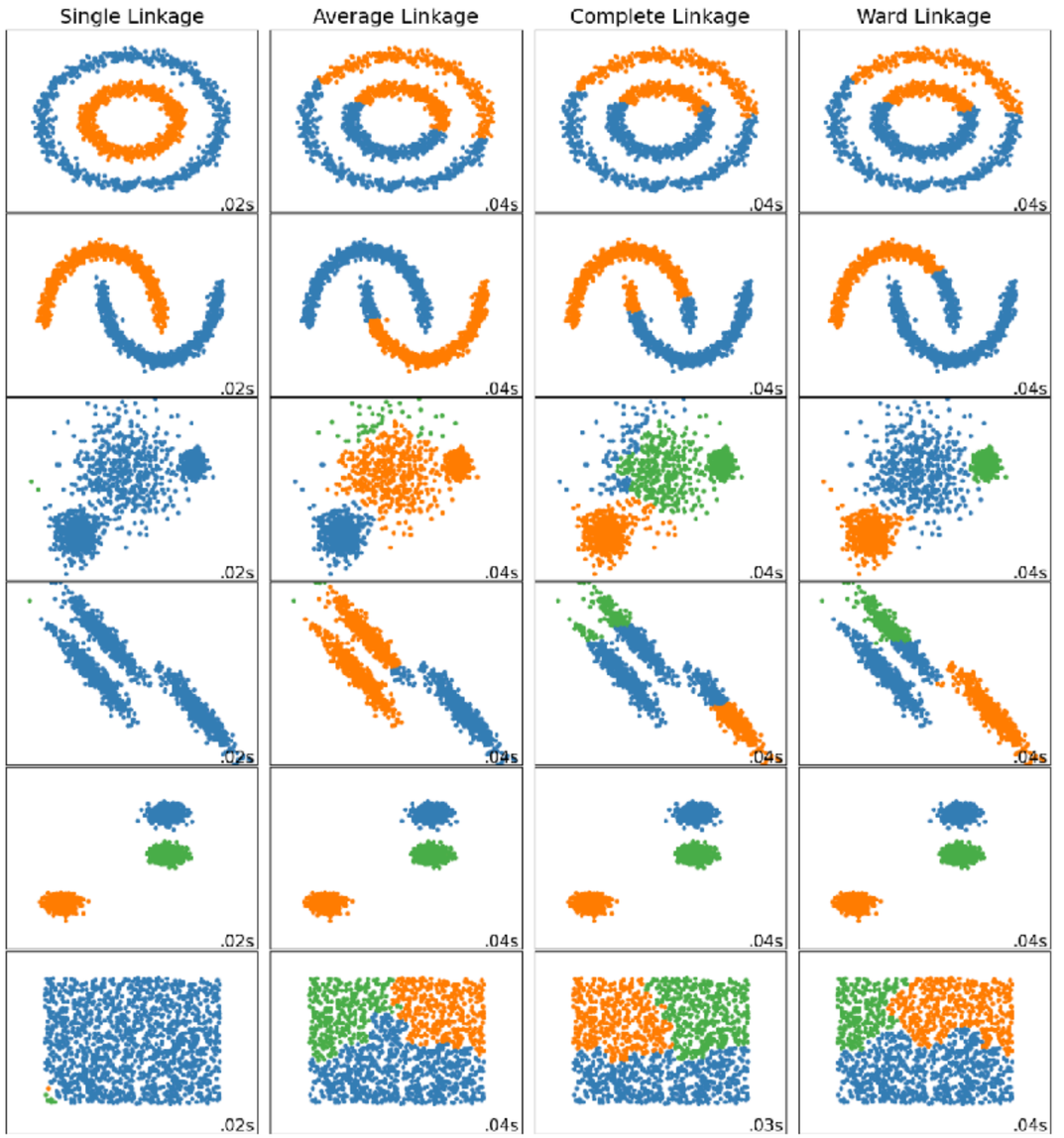
Agglomerative hierarchical clustering (AHC)

Ward minimizes the sum of squared differences within all clusters

Maximum or complete linkage minimizes the maximum distance between observations of pairs of clusters

Average linkage minimizes the average of the distances between all observations of pairs of cluster

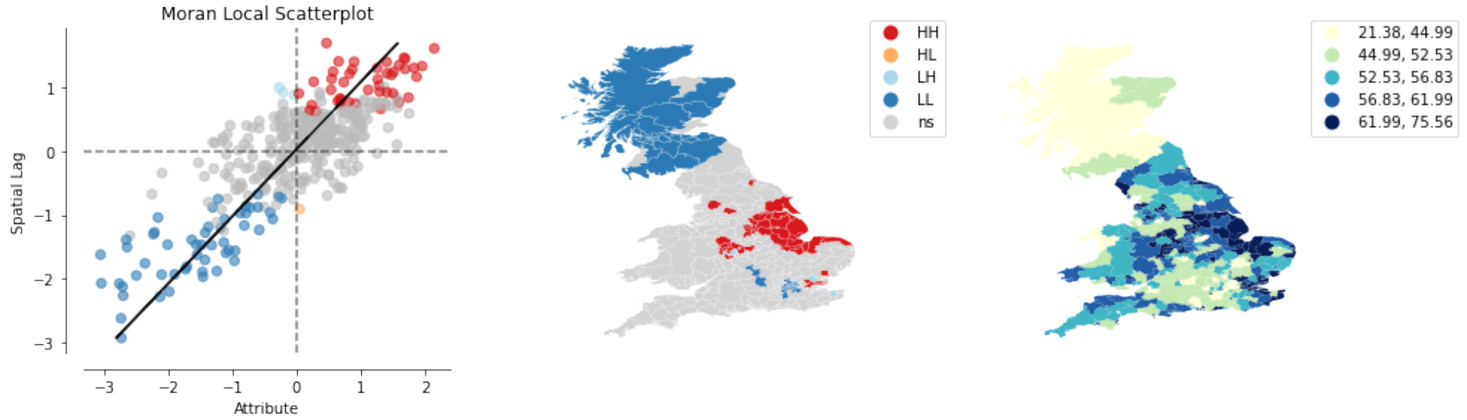
Single linkage minimizes the distance between the closest observations of pairs of clusters



Performance of spatial clustering is lowered by spatial constraints

Spatial coherence usually lowers ‘goodness of fit’

Use spatial autocorrelation to inform decision on spatial constraint



Comparing goodness of fit / feature coherence

Calinski-Harabasz Index

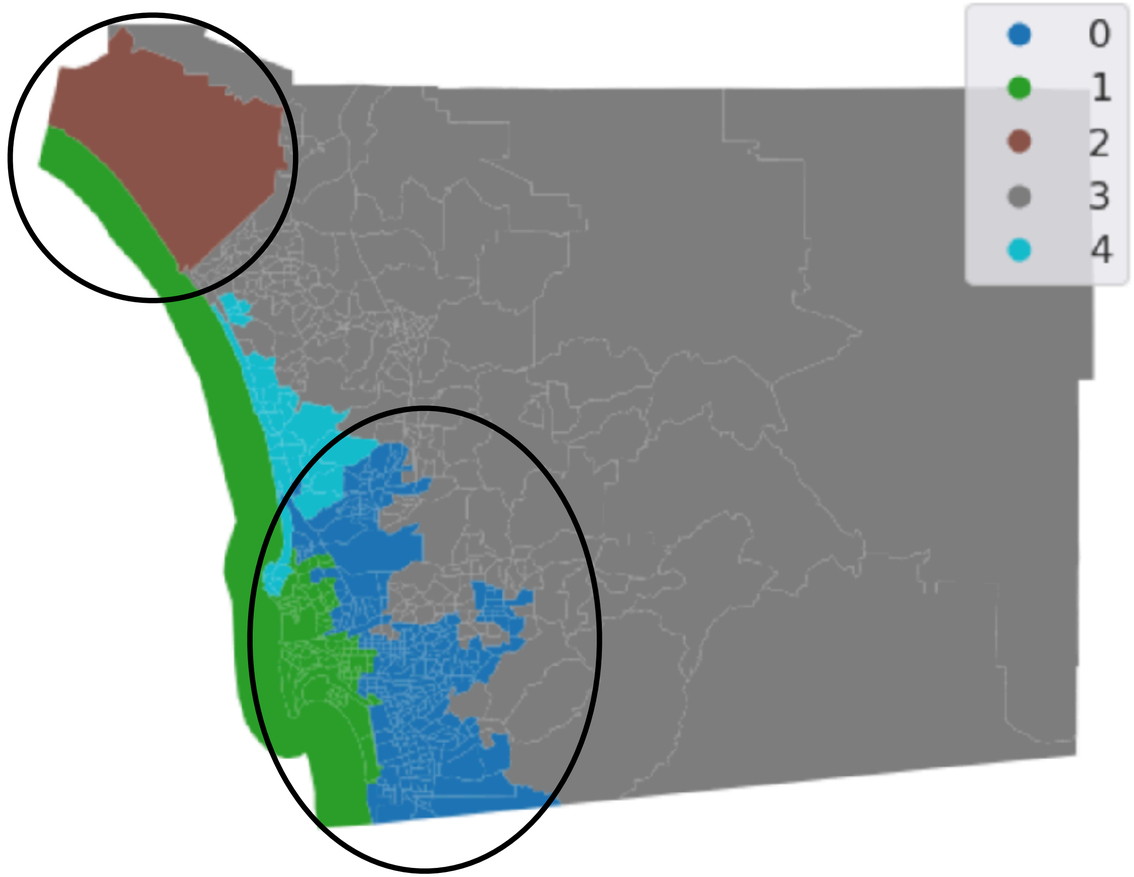
“the within-cluster variance divided by the between-cluster variance”

Silhouette Score

“the average standardized distance from each observation to its “next best fit” cluster”

Geographical coherence

Geographical coherence measures the spatial performance of regionalization



Geographical coherence

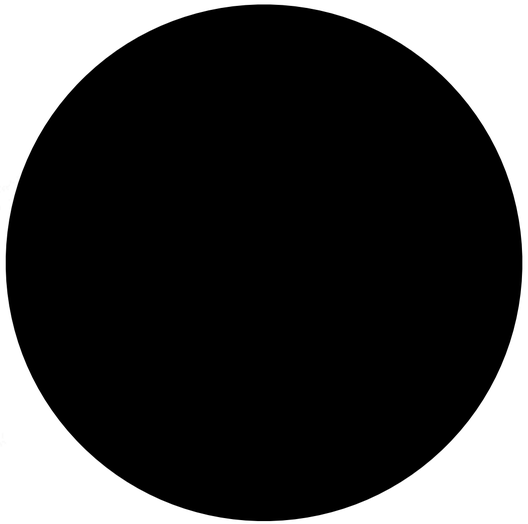
Isoperimetric quotient compares the area of the region to the area of a circle with the same perimeter as the region

Area of region

$$IPQ_i = \frac{A_i}{A_c} = \frac{4\pi A_i}{P_i^2}$$

Area of circle with same perimeter

Perimeter of region

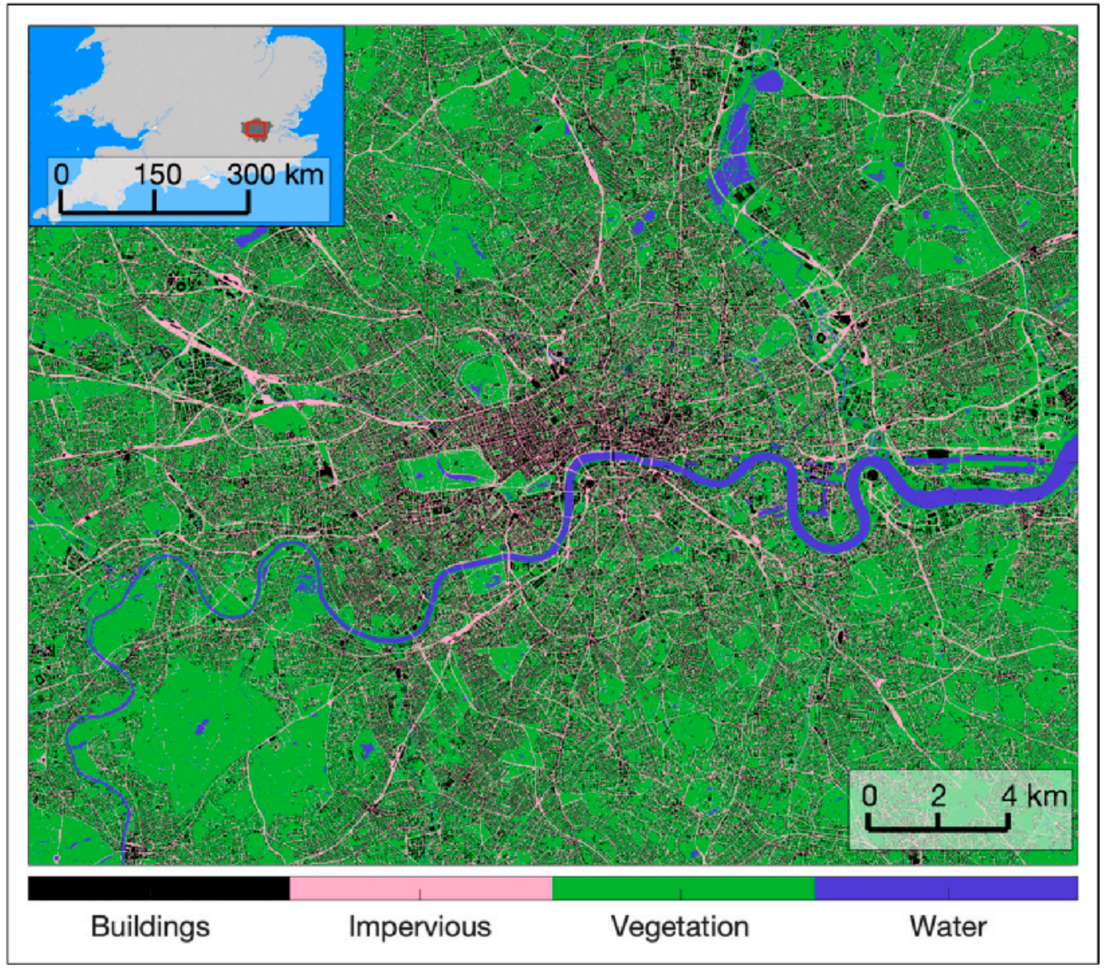


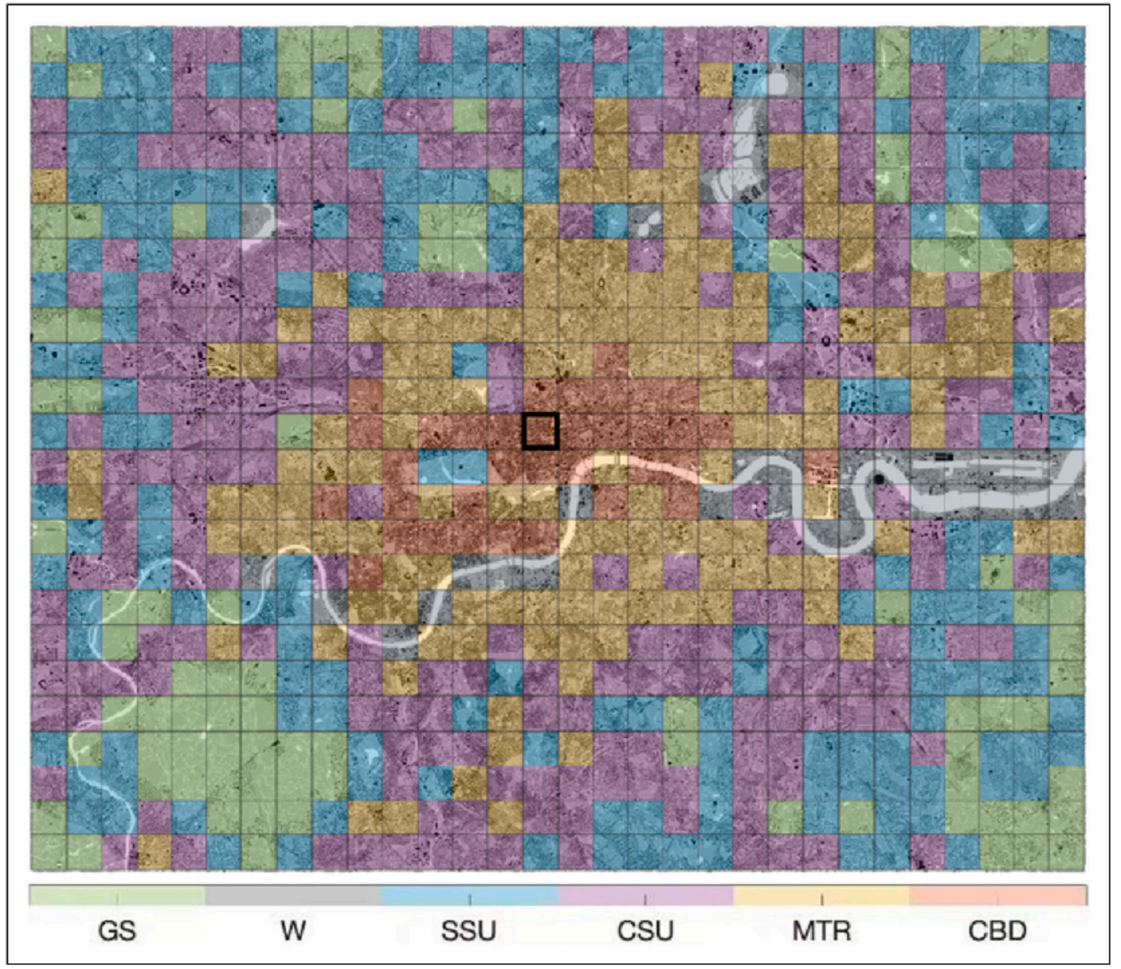
Isoperimetric quotient

IPQ closer to 0

IPQ closer to 1

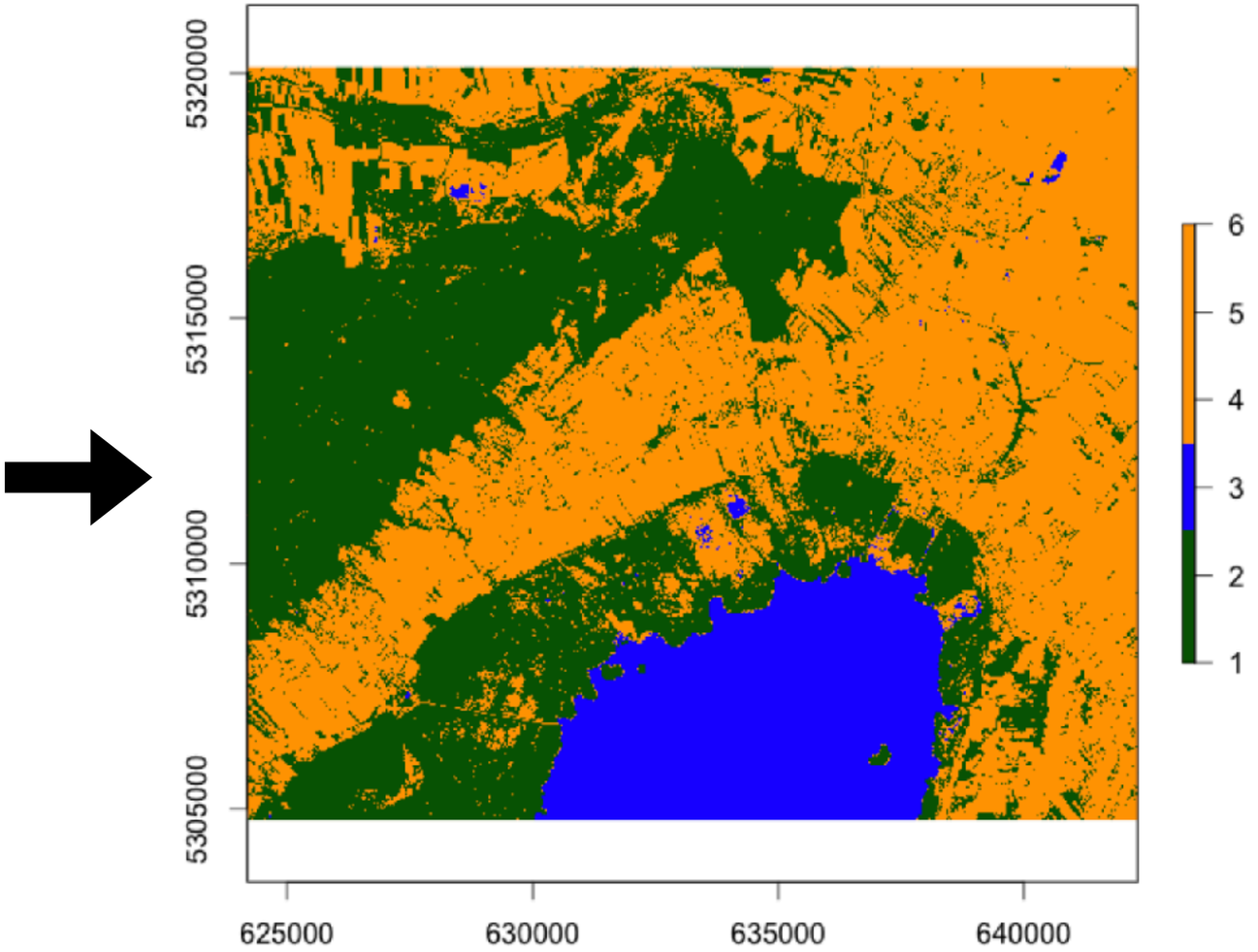
When do we use clustering in GDS?





Yu et al 2022

When do we use clustering in GDS?



https://www.gis-blog.com/unsupervised-kmeans-classification-of-satellite-imagery-using-r/

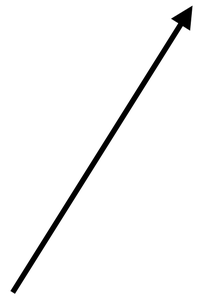
scikit-learn

```
kmeans5 = cluster.KMeans(n_clusters=6, random_state=12345)

k5cls = kmeans5.fit(data[variables])
```

```
sagg13 = cluster.AgglomerativeClustering(n_clusters=12, connectivity=w.sparse)

sagg13cls = sagg13.fit(data[cluster_cols_scaled])
```



spatial constraint

https://scikit-learn.org/stable/index.html

Jupyter

Sources and further materials for today's class

Geographic Data Science with Python

https://geographicdata.science/book/notebooks/10_clustering_and_regionalization.html

https://darribas.org/gds_course/content/bG/concepts_G.html

Next week: Point Pattern Analysis