Data Mining and Big Data Analytics

Spatial Data Mining

Class 10

Timur Naushirvanov
CEU

Outline

DATA MANAGEMENT AND ENGINEERING	
1	Introduction to data mining tasks and data types
2	Preprocessing and feature engineering: Data curation and filtering, imputation, scaling, dealing with categorical variables, features selection.
DATA MINING METHODS	
3	Basic classification methods: Decision trees, K-nearest neighbors, Naïve Bayes Classifier
4	Model evaluation: Generalization, overfitting and underfitting. Cross-validation. Model evaluation and comparison (e.g., metrics for classification, metrics for regression, confusion matrix, precision-recall curves, ROC curves).
5	Advanced classification methods: Support Vector Machine, Logistic regression, ANN
6	Hands-on session: Application of concepts on data and real-world situations.
7	Basic clustering methods: distance-based (separation, centroids, contiguity), density-based, partitional vs. hierarchical. Methods for centroid-based clustering (k-means), hierarchical clustering (single, complete and average linkage), density-based clustering (DBSCAN).
8	Outlier analysis: Extreme value analysis, Probabilistic methods, distance and density-based methods for outlier detection
9	Dimensionality reduction: Simple Value Decomposition, Principal Component Analysis, Embedding
SPECIFIC DATA MINING AREAS	
10	Spatial data mining: location inference, spatial demography inference, spatial trajectory reconstruction, learning from remotely sensed data
11	Graph data mining: network embedding, community detection methods
12	Final project presentation

Session Plan

- Introduction to Spatial Data Mining
  - Notebook
- Spatial Autocorrelation
- Spatial Clustering
  - Notebook
- Point Pattern Analysis
- Trajectory Analysis
  - Notebook

Menti

When you hear the term ‘spatial data mining’, what first comes to your mind?

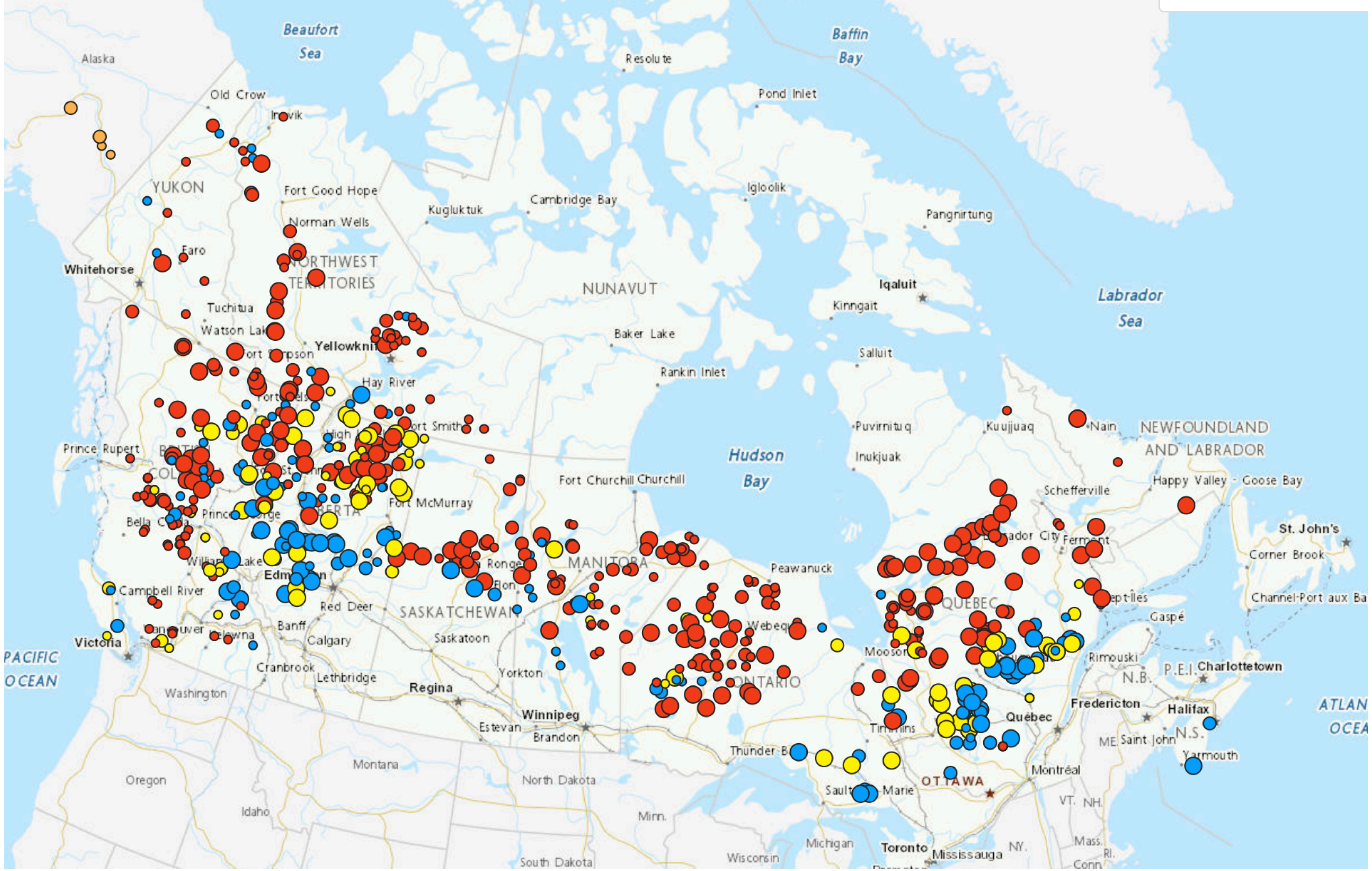
Go to menti.com

and use code: 8649 6752

Introduction to Spatial Data Mining

Mapping Wildfires in Canada

- Active Fires in Canada on 8 July 2023
- Data is collected from fire management agencies coordinated by the Canadian Interagency Forest Fire Centre (CIFFC) and Natural Resources Canada (NRCan)
- To assess burn areas: helicopter GPS flight, air photography, Landsat image classification



Source: Active Wildfires in Canada, ESRI

What is Spatial Data Mining?

Spatial Data Mining is a *non-trivial* search for *interesting* and *unexpected* spatial patterns.

Goals:

- Identifying spatial patterns
- Identifying spatial objects that are potential generators of spatial patterns
- Identifying information relevant for explaining the spatial pattern
- Presenting information in a way that is intuitive and supports further analysis

Source: Introduction to Spatial Data Mining, Stiftung Universität Hildesheim

Applications

- Meteorological Data
- Disease Outbreaks
- Mobile Objects
- Medical Diagnostics
- Earth Science
- Demographic Data

Applications

- Meteorological Data
  - Identifying patterns in weather data to predict the occurrence and movement of hurricanes, tornadoes, or other weather events.
- Mobile Objects
- Earth Science
- Disease Outbreaks
- Medical Diagnostics
- Demographic Data

Applications

- Meteorological Data
  - Identifying patterns in weather data to predict the occurrence and movement of hurricanes, tornadoes, or other weather events.
- Mobile Objects
- Earth Science
- Disease Outbreaks
  - Analysing environmental factors such as air quality or proximity to water sources to understand the spread of vector-borne diseases like malaria.
- Medical Diagnostics
- Demographic Data

Applications

- Meteorological Data
  - Identifying patterns in weather data to predict the occurrence and movement of hurricanes, tornadoes, or other weather events.
- Mobile Objects
  - Tracking and analysing the movement patterns of vehicles for optimising transportation routes or traffic management.
- Earth Science
  - Analysing satellite imagery to detect changes in land cover or land use over time, such as deforestation or urban sprawl.
- Disease Outbreaks
  - Analysing environmental factors such as air quality or proximity to water sources to understand the spread of vector-borne diseases like malaria.
- Medical Diagnostics
  - Using spatial data mining to identify patterns in medical imaging data for early detection of diseases such as cancer or Alzheimer's disease.
- Demographic Data
  - Studying spatial variations in socioeconomic indicators such as income levels or educational attainment to identify areas in need of targeted social programs.

Types of Spatial Data

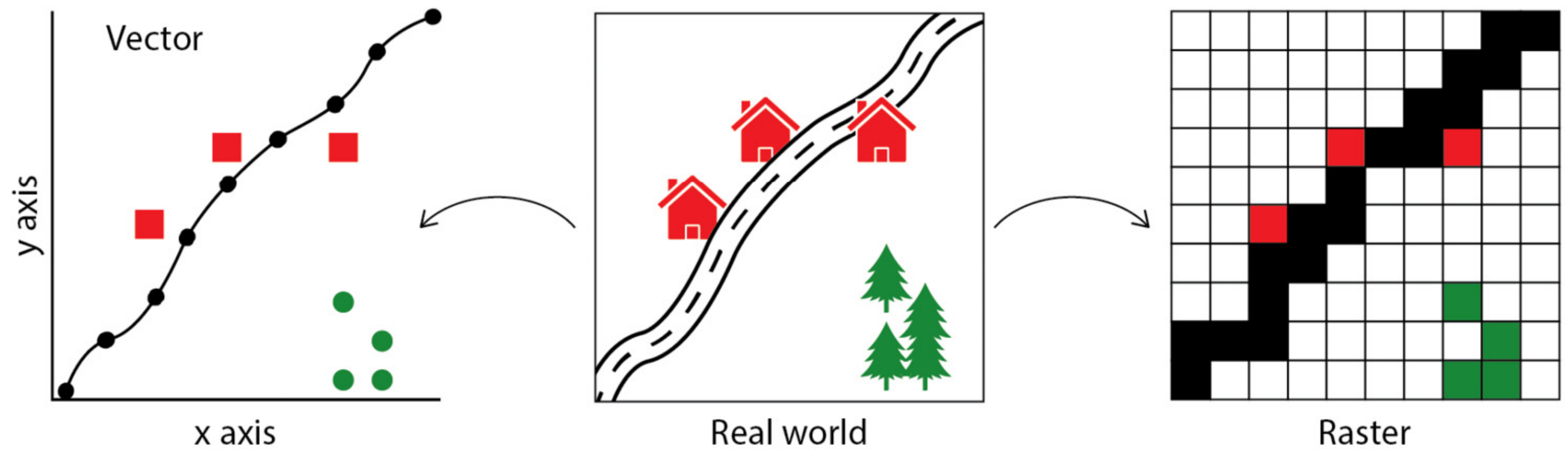
Spatial data is data that have some form of *spatial* or *geographic reference* that enables them to be *located in two or three dimensional space*.

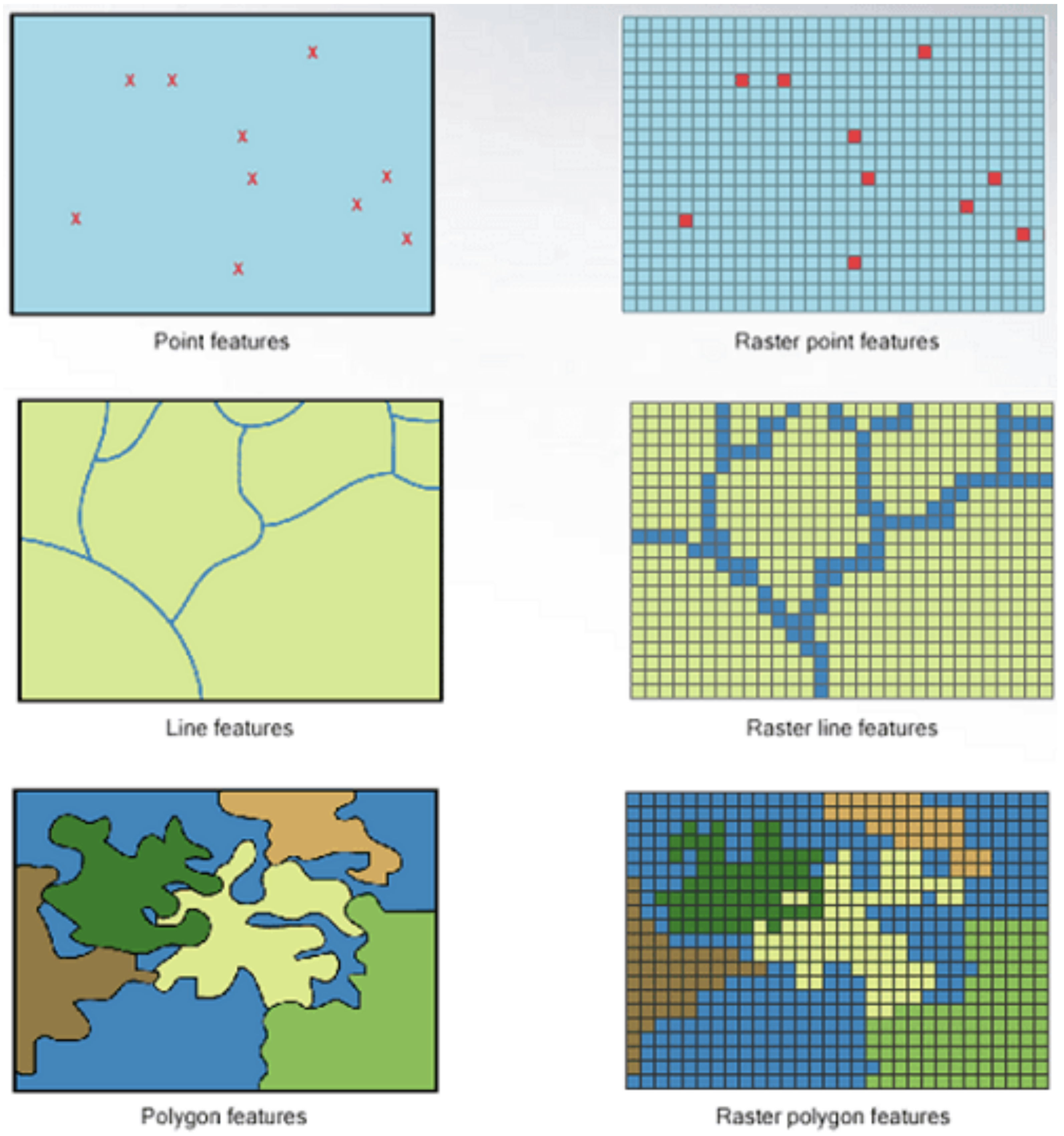
GIS - Geographic Information Systems - is a system to represent and analyse spatial data.

1. **Feature (or Vector) Data**
    - Describes the features of geographic locations through the use of discrete geometries: Point, Line, Polygon.
2. **Coverage (or Raster) Data**
    - Encodes the world as a continuous surface represented by a grid. Each values of a grid can be either a continuous value or a categorical classification.
        - Satellite images, altitude maps, etc.

Sources: GeoPandas Tutorial: An Introduction to Geospatial Analysis;
Introduction to Spatial Data Mining, Stiftung Universität Hildesheim

Types of Spatial Data





Sources: David S. Jordan, Applied Geospatial Data Science with Python, Chapter 1;
[GeoPandas Tutorial: An Introduction to Geospatial Analysis]

Notebook 1

Introduction to GeoPandas

Spatial Data Formats

CRS (Coordinate Reference System)

Spatial Data Mining:

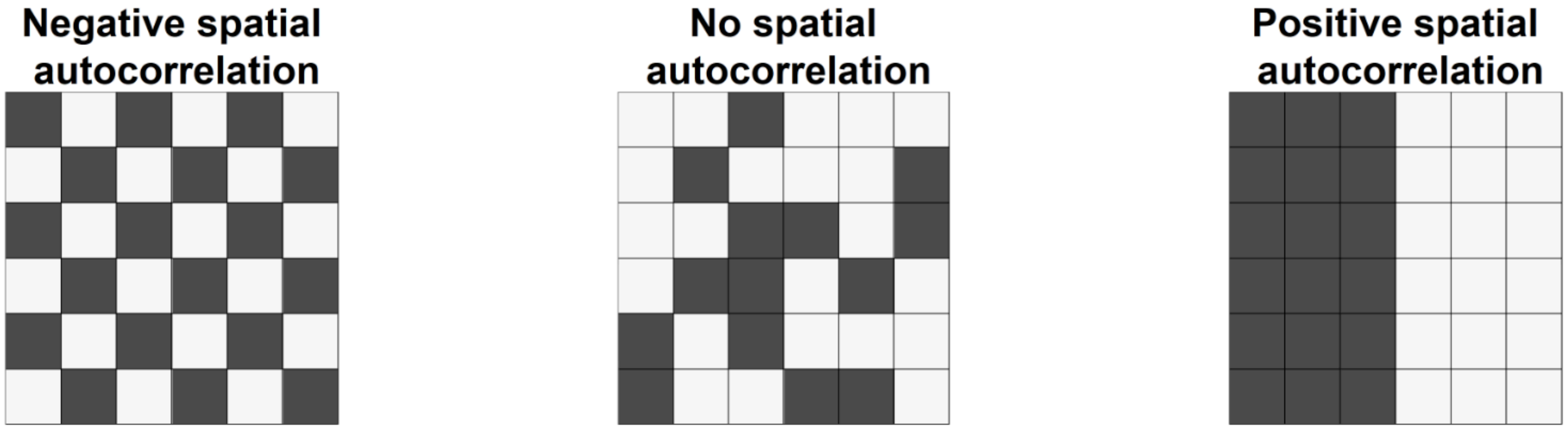
Spatial Autocorrelation

Spatial Clustering

Point Pattern Analysis

Spatial Autocorrelation

- **Spatial Autocorrelation** describes the degree to which the similarity in values between observations is correlated to the similarity in locations of such observations.

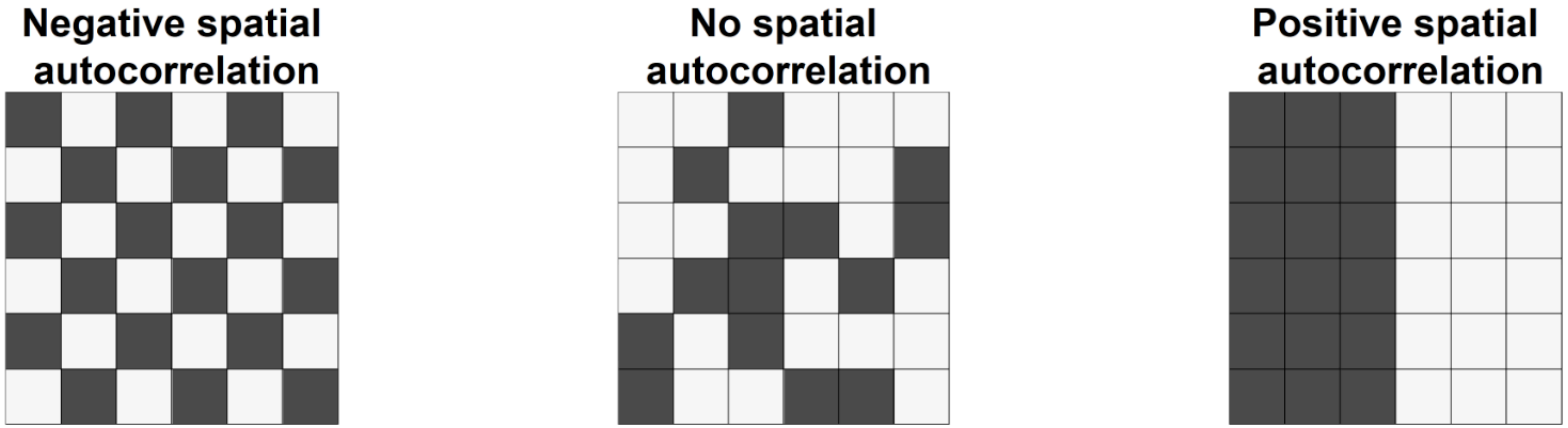


- Alternative explanation: it is the degree of information contained in the value of a variable at a given location about the value of that same variable in other locations.

Source: Geographic Data Science with Python

Spatial Autocorrelation

- **Spatial Autocorrelation** describes the degree to which the similarity in values between observations is correlated to the similarity in locations of such observations.



similar values tend to be located away from each other

phenomena of spatial competitions: gyms/stores of different brands

spatial randomness

similarity and geographical closeness go hand-in-hand

spatial segregation in cities

Source: Geographic Data Science with Python

Spatial Autocorrelation

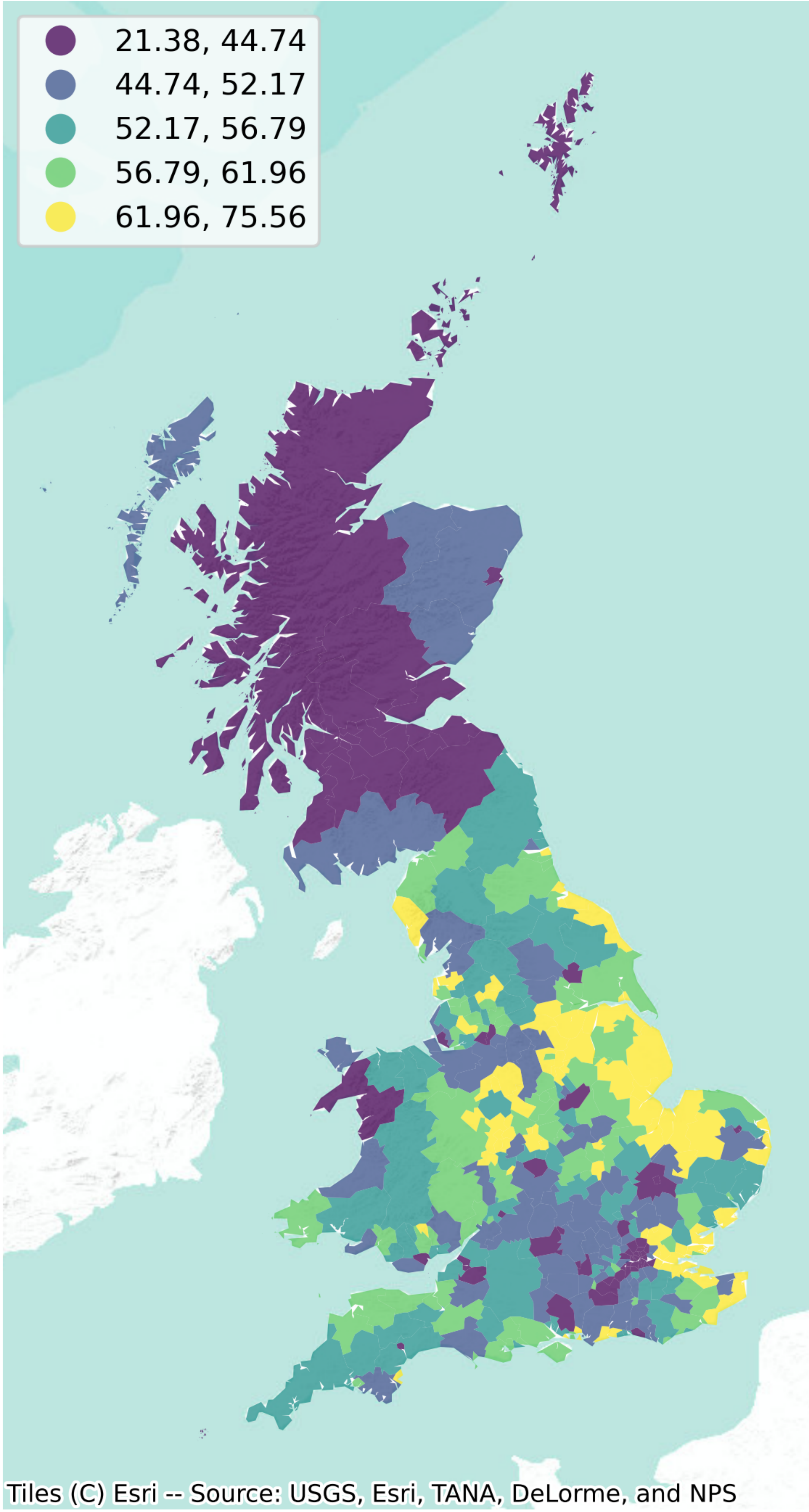
- **Spatial Autocorrelation** describes the degree to which the similarity in values between observations is correlated to the similarity in locations of such observations.

- Two types:

Global	**Local**
Helps to see the overall trend that the location of values follows.	Focuses on on the relationships between *each* observation and its surroundings.
Makes possible statements about the degree of *clustering* in the dataset.	Makes possible statements about the degree of *clustering* in the dataset.
Are similar values closer to other similar values than we would expect from pure chance?	*What localised areas exhibit significant concentrations of high or low temperature anomalies compared to their immediate surroundings?*

Source: Geographic Data Science with Python

Spatial Autocorrelation: Case



- **UK 2016 Brexit Referendum**
- On the map - percentage of people who voted ‘Leave’ (divided in 5 quantiles)
- Spatial weights used: eight nearest neighbours

Source: Geographic Data Science with Python

Spatial Autocorrelation

Spatial Weights is a construct used to represent geographic relationships between the observational units in a spatially referenced dataset. It is the notion of geographical proximity or connectedness.

1. **Contiguity weights**
    - A pair of spatial objects share a common border.
2. **Distance-based weights**
    - kNN, Great circle…
3. **Block weights**
    - Membership in a geographic group defines the neighbour relationships.

Source: Geographic Data Science with Python

Spatial Autocorrelation: Case

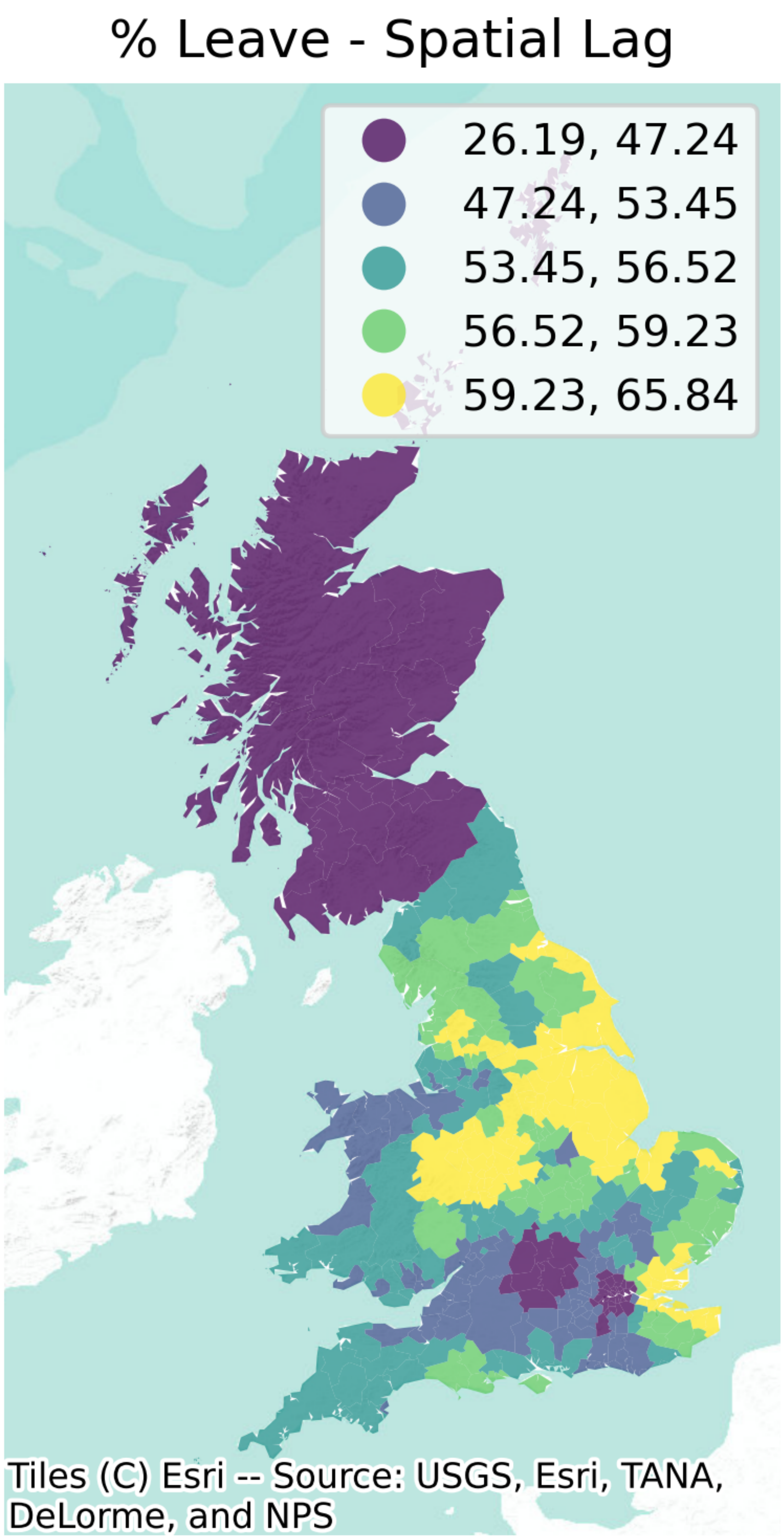
Spatial lag

$$Y_{sl} = \mathbf{WY}$$

$$y_{sl-i} = \sum_j w_{ij} y_j$$

	Pct_Leave	Pct_Leave_lag
Liverpool	41.81	54.61375
Midlothian	37.94	38.01875

- If **W** is binary:
  - Spatial lag becomes a sum of the values of *i*'s neighbours
- If **W** is row-standardised:
  - Spatial lag becomes the average value of **Y** in the neighbourhood of each observation *i*

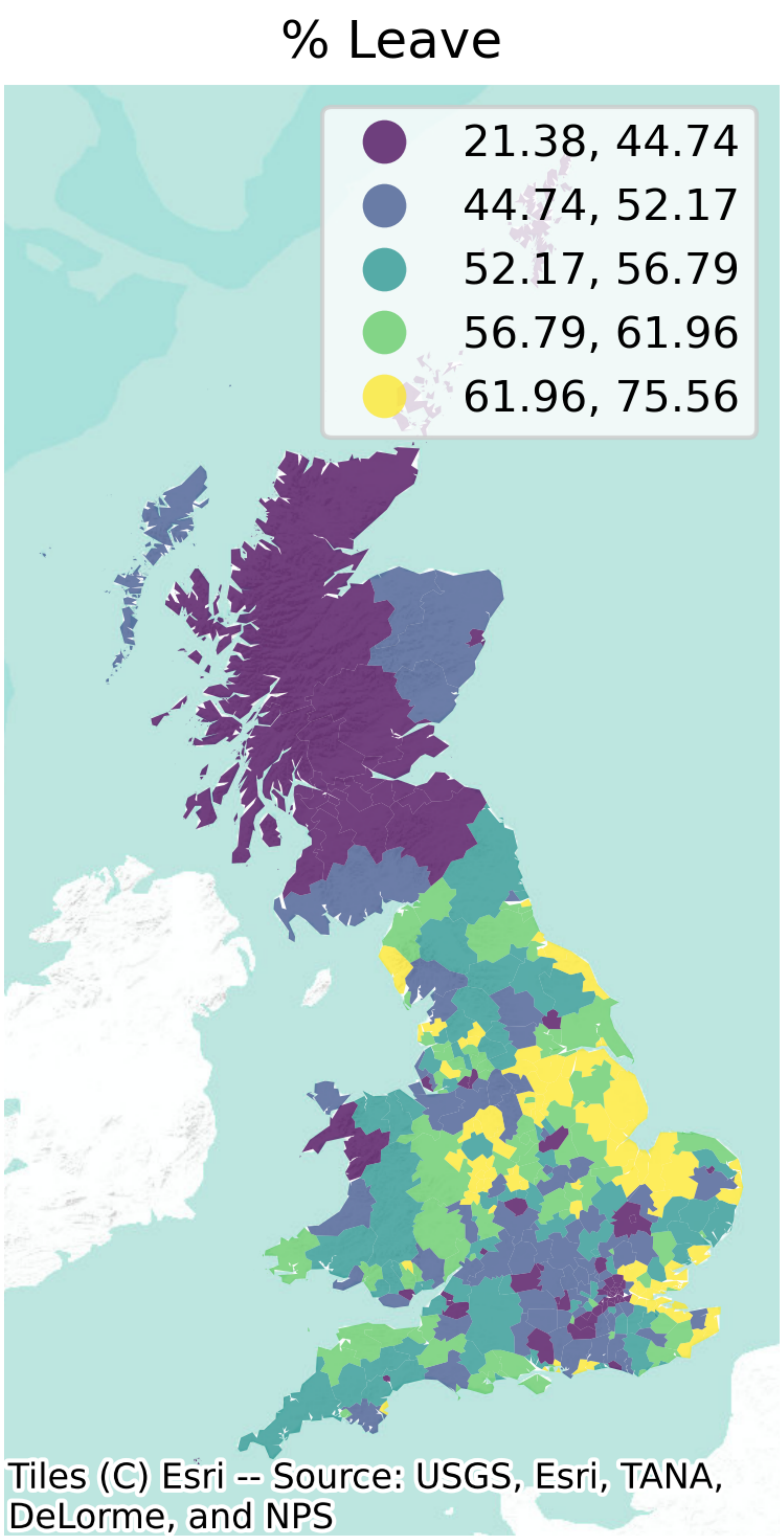


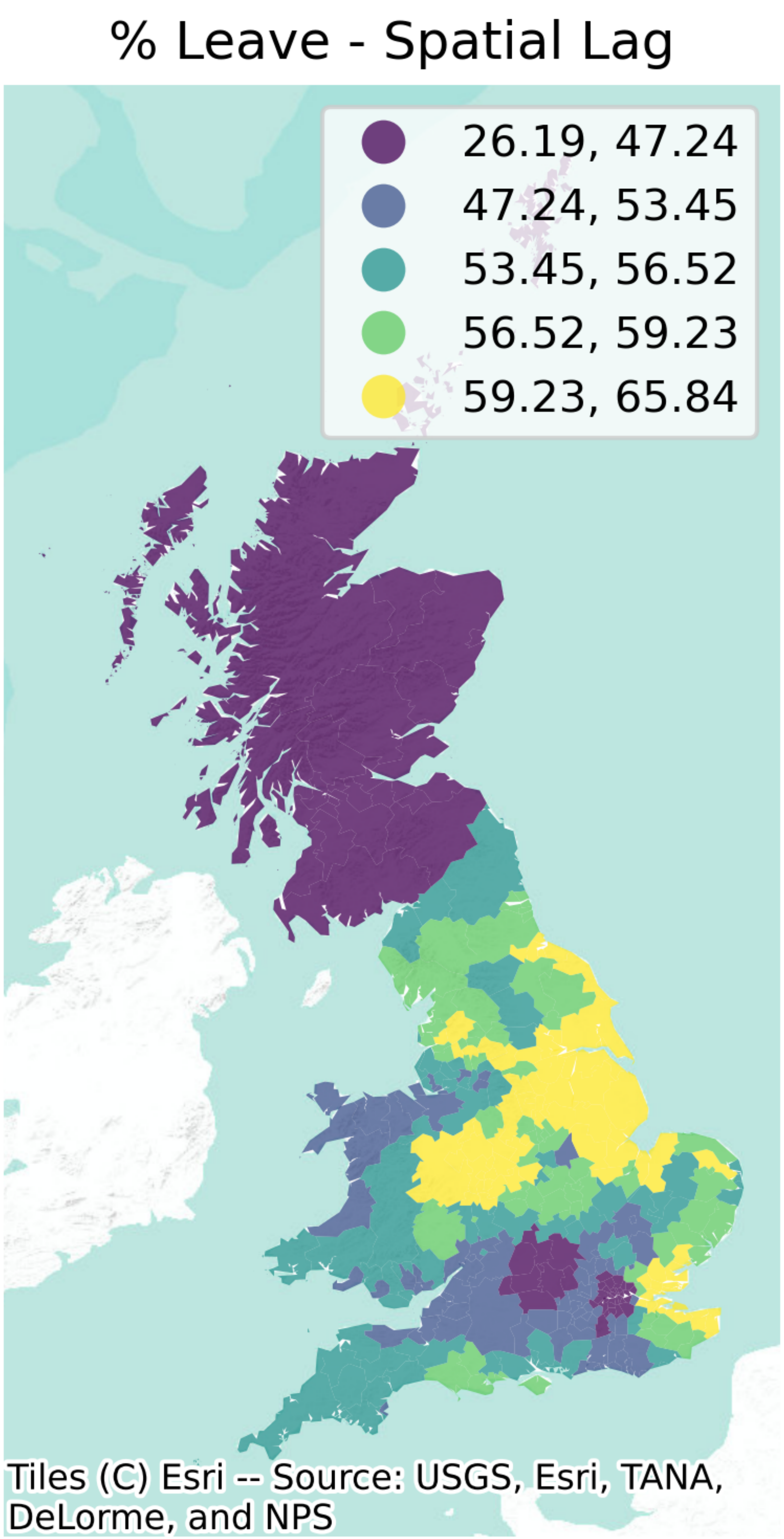
Source: Geographic Data Science with Python

Spatial Autocorrelation: Case

The **spatial lag** can smooth out the differences between nearby observations.

Using **spatial lag**, we can begin to relate the behaviour of a variable at a given location to its pattern in the immediate neighbourhood.





Source: Geographic Data Science with Python

Spatial Autocorrelation

- **Spatial Autocorrelation** describes the degree to which the similarity in values between observations is correlated to the similarity in locations of such observations.

- Two possible indices to compute spatial autocorrelation:

Moran's *I*

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (Y_i - \bar{Y})^2},$$

n - number of regions (spatial units)

Y_i - the observed variable in region *i*

$\bar{Y}$ - the mean of Y

w_{ij} - spatial weights denoting spatial proximity

Geary's *C*

$$C = \frac{(N-1) \sum_i \sum_j w_{ij}(x_i - x_j)^2}{2S_0 \sum_i (x_i - \bar{x})^2}$$

N - number of regions (spatial units)

x_i - the observed variable in region *i*

$\bar{x}$ - the mean of x

w_{ij} - spatial weights denoting spatial proximity

S_0 - sum of all w

Source: Spatial Statistics for Data Science: Theory and Practice with R

Spatial Autocorrelation: Moran's *I*

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

- Under the null hypothesis of no spatial autocorrelation, observations Y_i are independent identically distributed, and *I* is asymptotically normally distributed with mean and variance equal to:

$$E[I] = \frac{-1}{n-1} \qquad Var[I] = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}$$

$$S_0 = \sum_{i \neq j} w_{ij},\ S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2 \text{ and } S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik} \right)^2$$

- Moran's *I* values usually range from –1 to 1. If they are significantly above E[*I*], it indicates positive spatial correlation or clustering.

Source: Spatial Statistics for Data Science: Theory and Practice with R

Spatial Autocorrelation: Moran's *I*

- When the number of regions is sufficiently large, *I* has a normal distribution and we can assess whether any given pattern deviates significantly from a random pattern by comparing the z-score to the standard normal distribution.

$$z = \frac{I - E(I)}{Var(I)^{1/2}}$$

- Alternatively, z-score can be compared to the values we get after using Monte Carlo randomisation.
- MC randomisation creates random patterns by reassigning the observed values among the areas and calculates the Moran's *I* for each of the patterns, providing a randomisation distribution for the Moran's *I*.

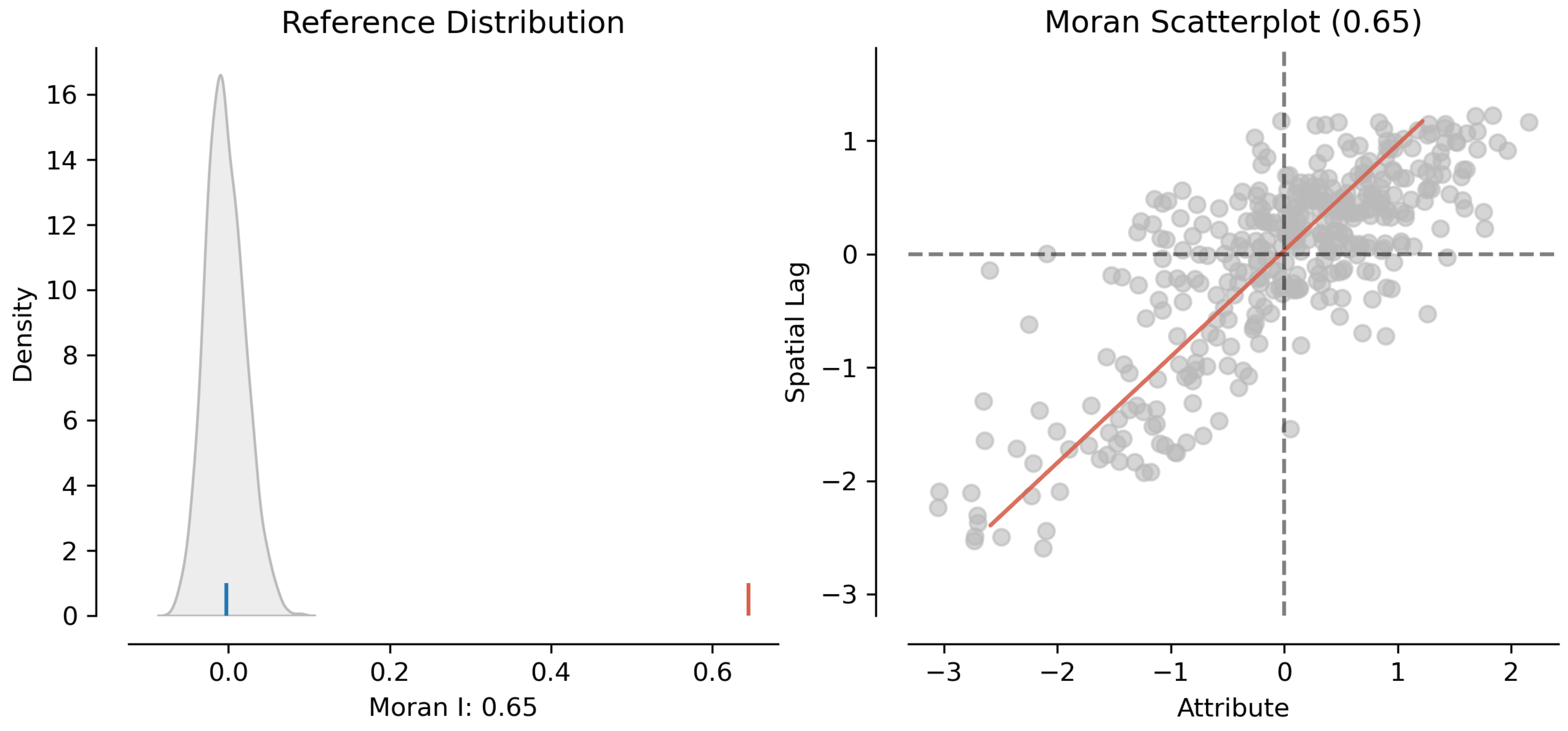
Source: Spatial Statistics for Data Science: Theory and Practice with R

Spatial Autocorrelation: Case

The relationship between the standardised "Leave" voting percentage in a local authority and its spatial lag (the average standardised density of the percent Leave vote in the neighbourhood of each observation).

A positive relationship indicates the presence of positive autocorrelation.

moran.I = 0.6455
moran.p_sim = 0.001
-> small enough p-values allows to reject the hypothesis that the map is random



Source: Spatial Statistics for Data Science: Theory and Practice with R

Spatial Autocorrelation: Geary's C

$$C = \frac{(N-1)\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2S_0 \sum_i (x_i - \bar{x})^2}$$

- The value of Geary's *C* lies between 0 and some unspecified value greater than 1.
- Values significantly lower than 1 demonstrate increasing positive spatial autocorrelation, whilst values significantly higher than 1 illustrate increasing negative spatial autocorrelation.
- Geary's *C* is inversely related to Moran's *I*, but not identical. Geary's *C* uses the sum of squared distances, whereas Moran's *I* uses standardised spatial covariance. By using squared distances, Geary's *C* is less sensitive to linear associations and may pickup autocorrelation where Moran's *I* may not.

Sources: Spatial Statistics for Data Science: Theory and Practice with R; Geary's C | Wikipedia

Spatial Autocorrelation: LISA

- Global Moran's *I* provides an index to assess the spatial autocorrelation for the whole study region; it can tell us whether values in our map *cluster* together (or disperse) overall, but it will not inform us about where specific *clusters* (or outliers) are.
- Alternatively, we can have a local measure of similarity between each area's value and those of nearby areas - Local Indicators of Spatial Association (**LISA**).

$$I_i = \frac{n(Y_i - \bar{Y})}{\sum_j (Y_j - \bar{Y})^2} \sum_j w_{ij}(Y_j - \bar{Y}) \qquad I = \frac{1}{\sum_{i \neq j} w_{ij}} \sum_i I_i$$

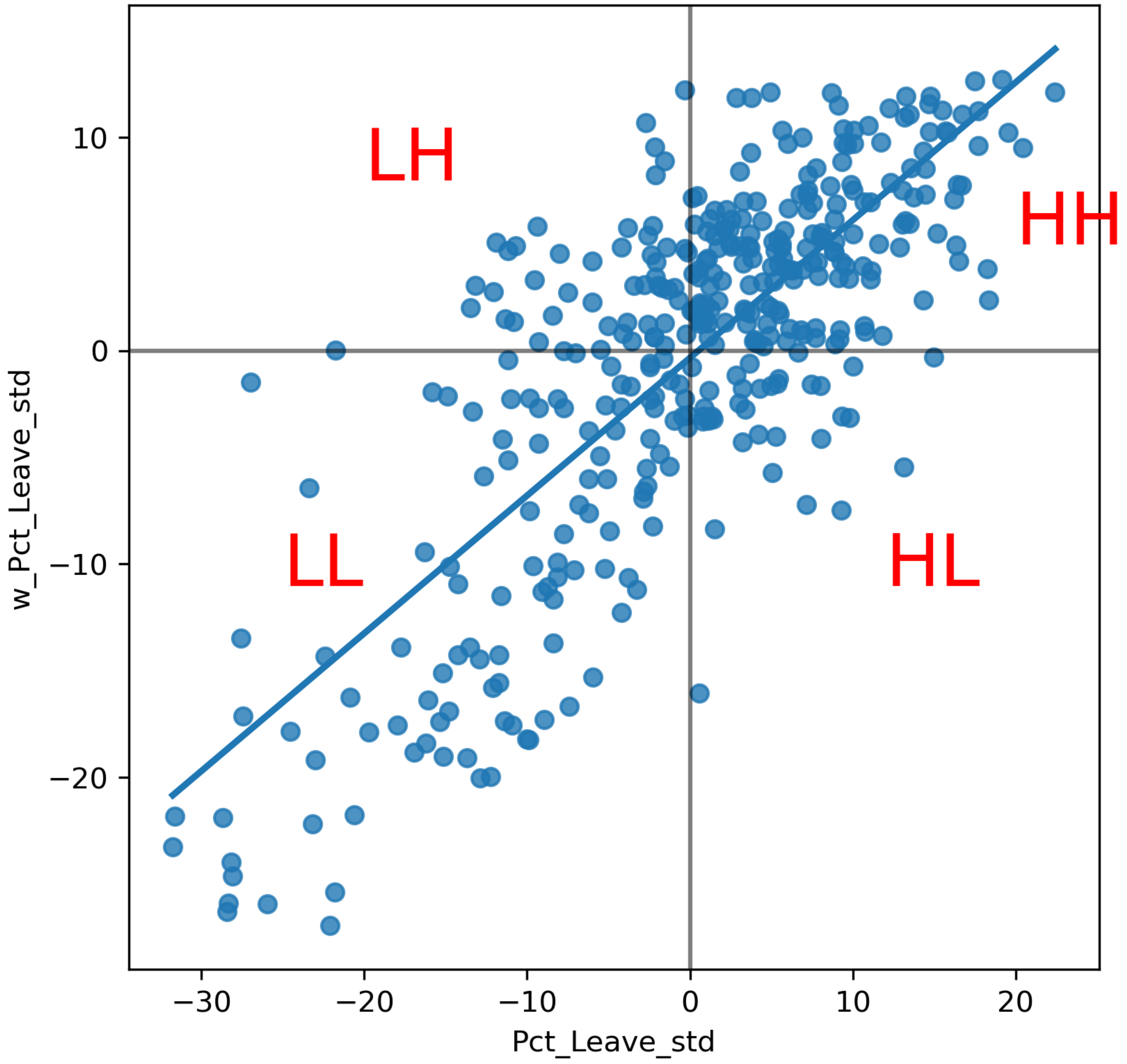
- The values of the LISAs are mapped to indicate the location of areas with comparatively high or low local association with neighbouring areas.
- A high value for I_i suggests that the area is surrounded by areas with similar values.

Source: Spatial Statistics for Data Science: Theory and Practice with R

Spatial Autocorrelation: Case

Divide into **quadrants** with each capturing a situation based on whether a given area displays a value above the mean (high) or below (low) in either the original variable or its spatial lag.

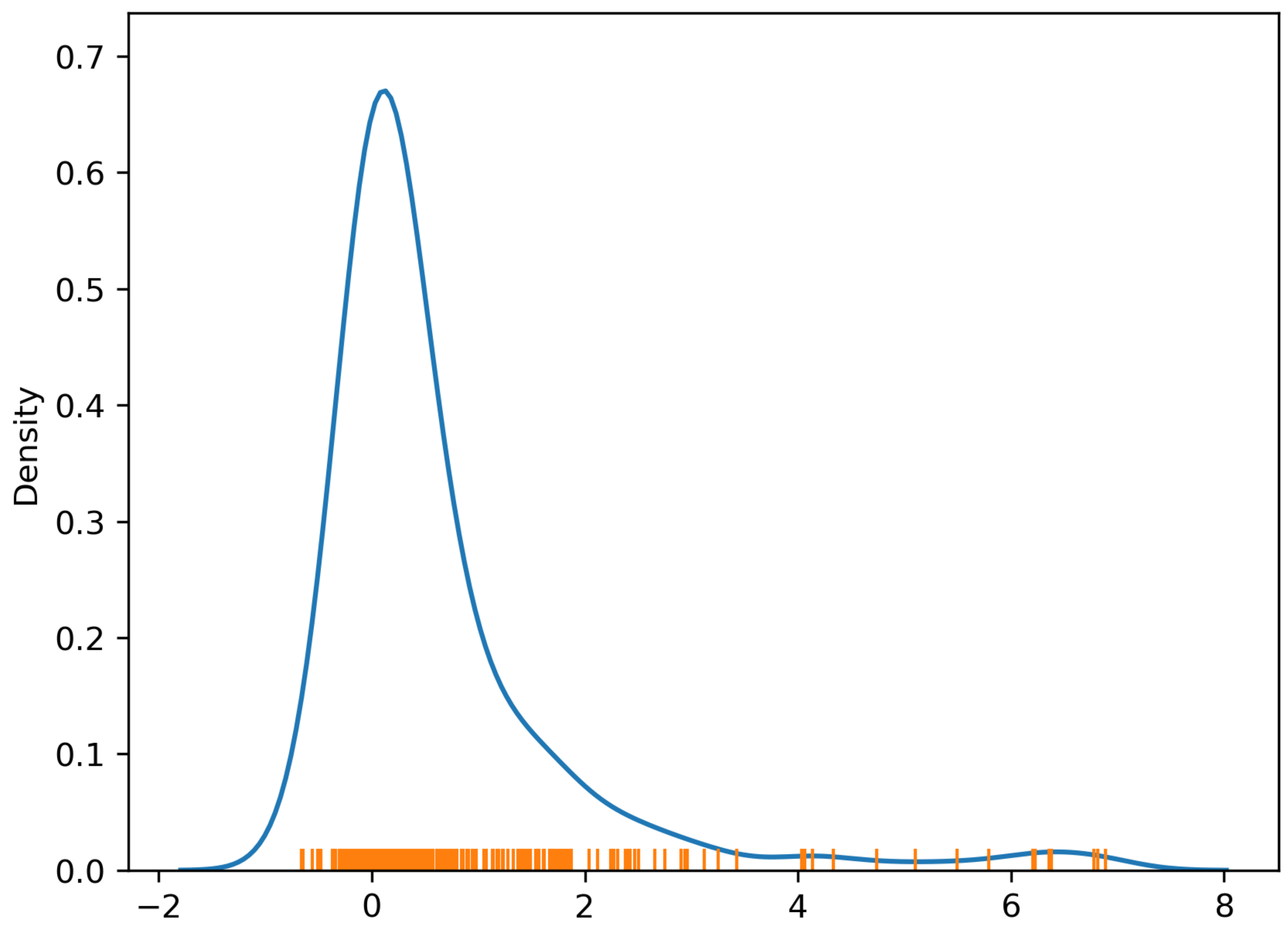
The **core idea**: identify cases in which the value of an observation and the average of its surroundings is either more similar or dissimilar, compared to the pure chance.



Source: Geographic Data Science with Python

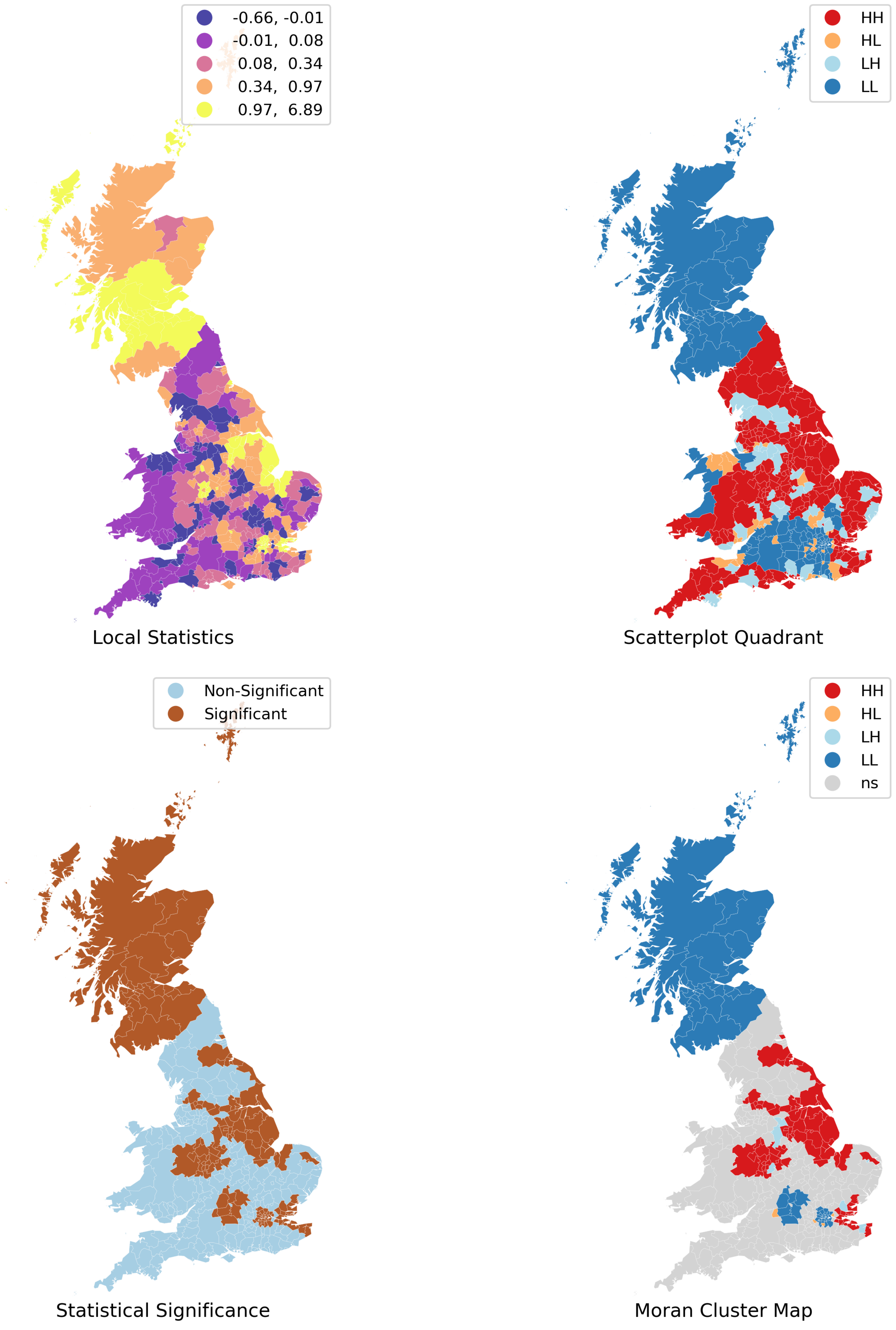
Spatial Autocorrelation: Case

The distribution of local Moran's I



Skewed - due to the dominance of positive forms of spatial association

Important to keep in mind: cannot differentiate between HH and LL, or between HL and LH



Source: [Geographic Data Science with Python]

Spatial Clustering

Spatial Autocorrelation

- Statistical measure that quantifies the degree of similarity between observations at different locations in space. It examines whether there is a relationship between the values of a variable at one location and the values at nearby locations.

Spatial Clustering

- Grouping of similar observations or values together within a geographic area. These clusters can be identified visually or through statistical analysis.
- Region can be perceived as a cluster in for spatial data (but geographically consistent).

Source: Spatial Clustering by Luc Anselin

Spatial Clustering

Clustering on Geographical Coordinates

- Results in clusters as *regions* in space.
- Critical that the geographical coordinates are projected.
- For most methods, the clusters will tend to result in fairly compact regions (Voronoi polygons, etc.)

Including Geographical Coordinates in the Feature Set

- No guarantee that resulting clusters are spatially contiguous (and not designed to be).
- One solution: include geometric centroids as part of clustering (projected!). But still does not guarantee contiguity.

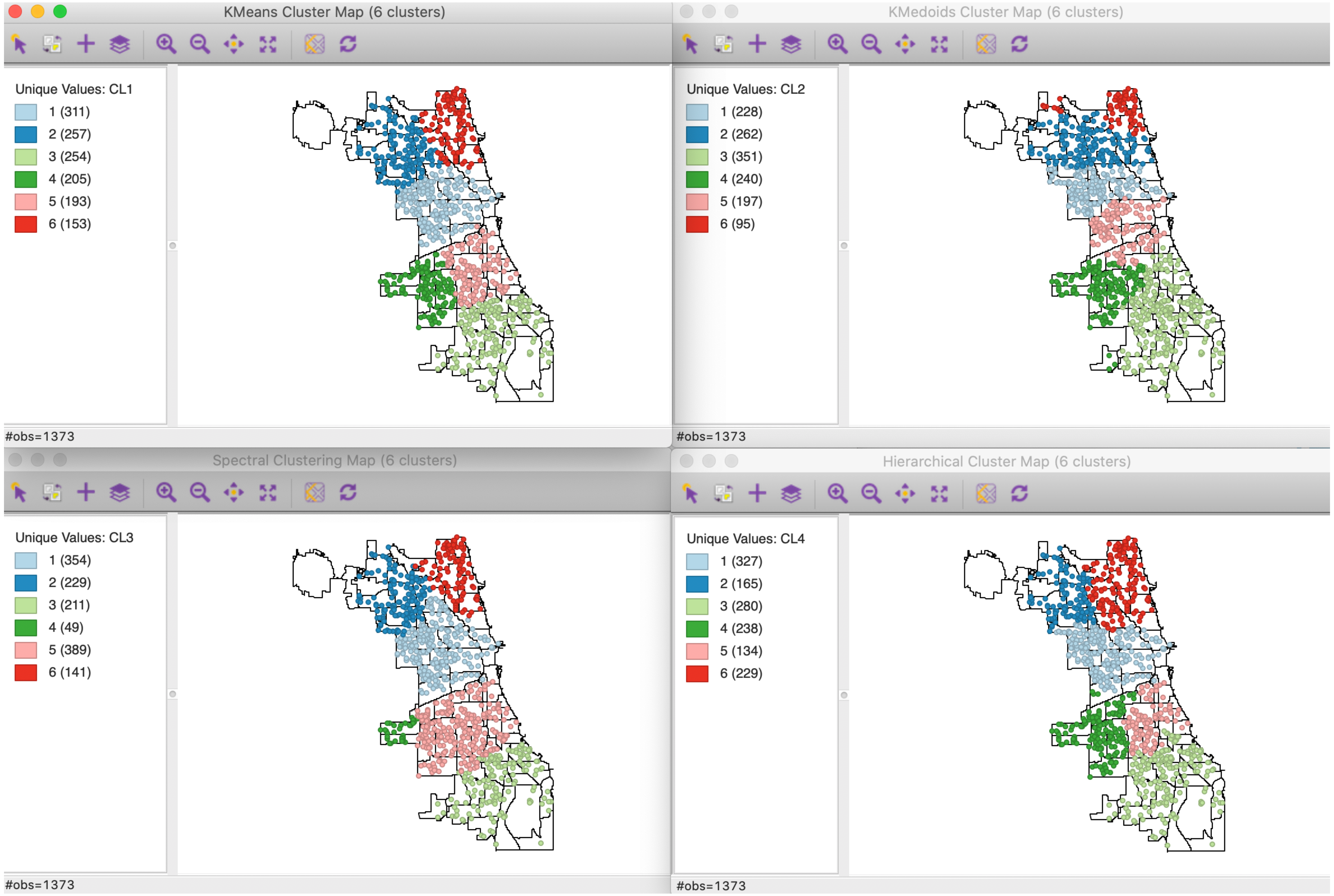
Weighted Optimisation of Attribute and Geographical Similarity

- Two functions: one is focused on the similarity of the regular attributes, the other on the similarity of the geometric centroids.
- A weight changes the relative importance of each objective.

Source: Spatial Clustering by Luc Anselin

Spatial Clustering: On Geographical Coordinates

1373 points against the Chicago community area boundaries



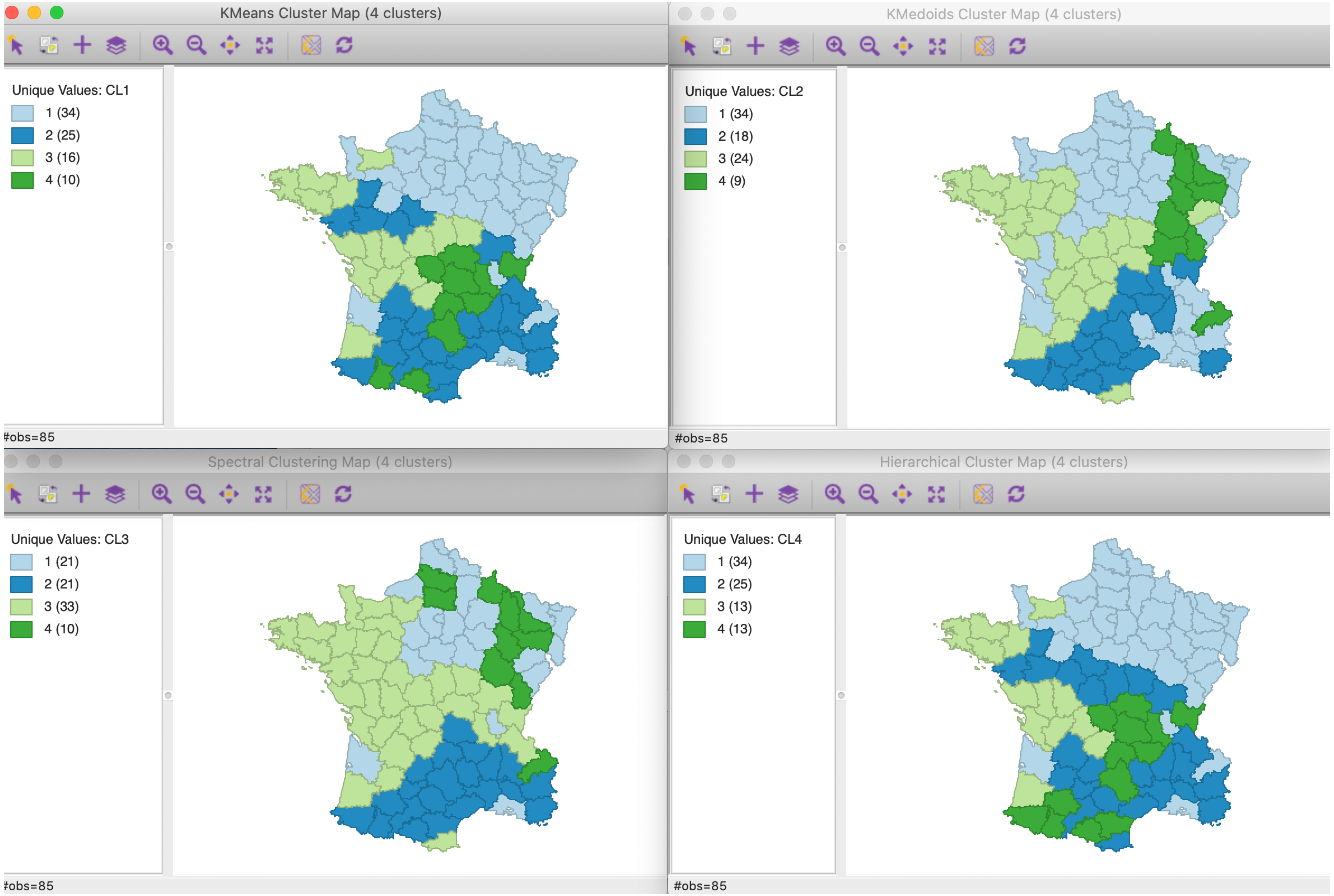
Source: Spatial Clustering by Luc Anselin

Spatial Clustering: Only Features

Based on 6 features: **Crm_prs**, **Crm_prp**, **Literacy**, **Donations**, **Infants** and **Suicides**

results differ by method

geographic grouping is far from being contagious



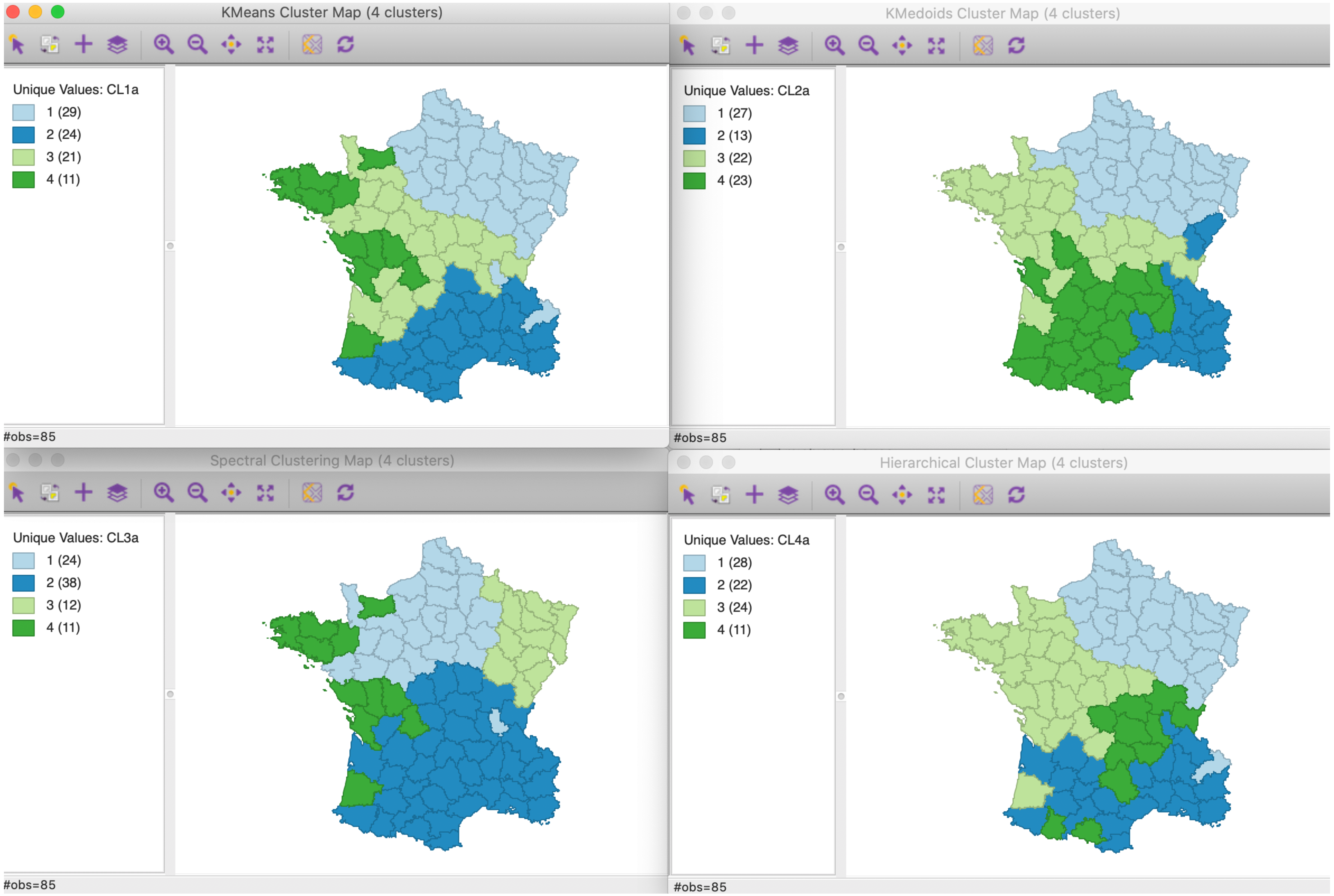
Source: Spatial Clustering by Luc Anselin

Spatial Clustering: Geo Centroids as Features

Adding coordinates of the centroids

results differ by method

geographic grouping is not perfectly contagious but spatially more structured



Source: Spatial Clustering by Luc Anselin

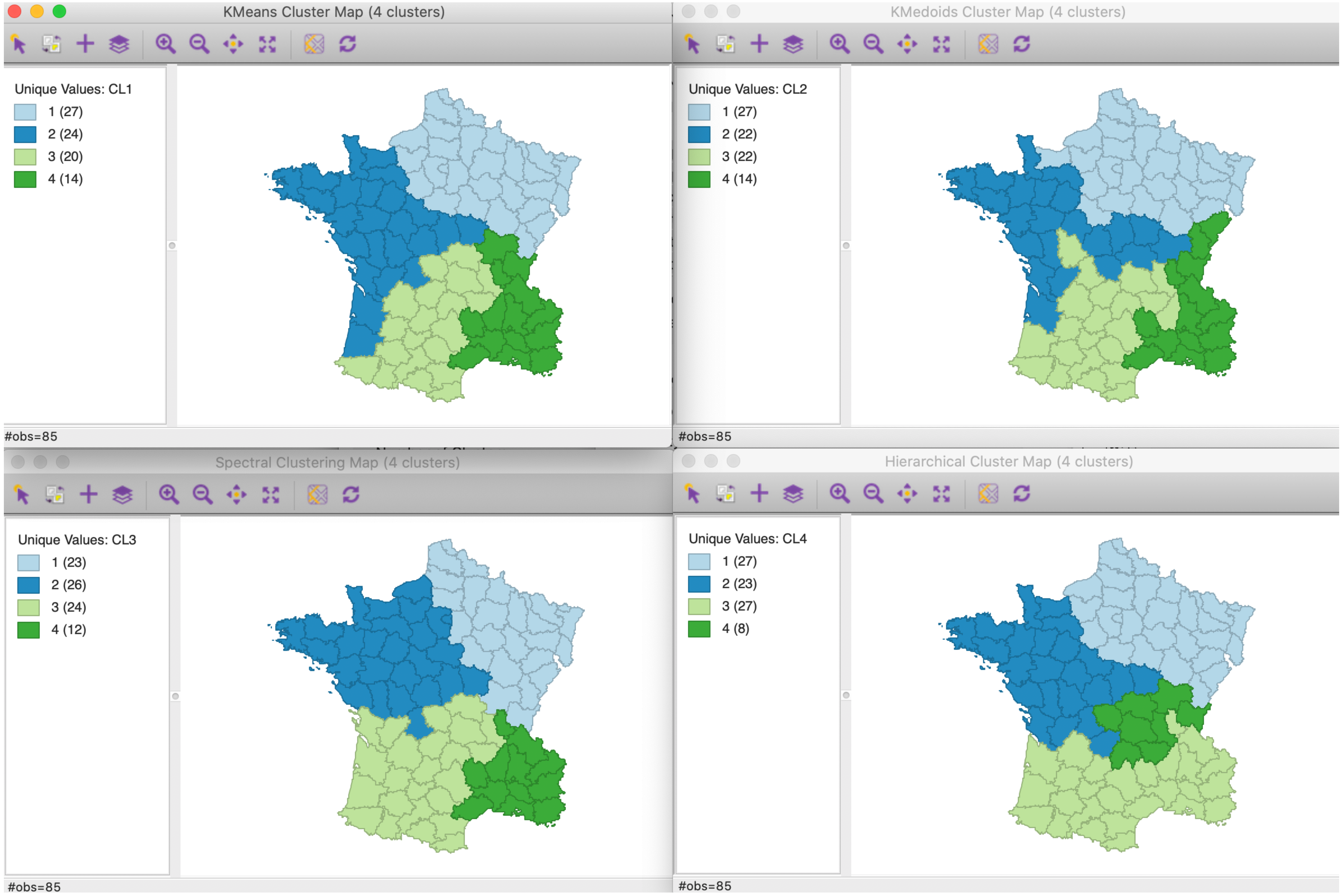
Spatial Clustering: Weighted Optimisation

coordinate variables are treated separately from the regular attributes

Example:
w1 - for geographic
w2 - for regular
w1 + w2 = 1

results yield more contiguity

illustrates the trade offs between attribute and locational similarity



Source: Spatial Clustering by Luc Anselin

Notebook 2

Spatial Autocorrelation

Spatial Clustering

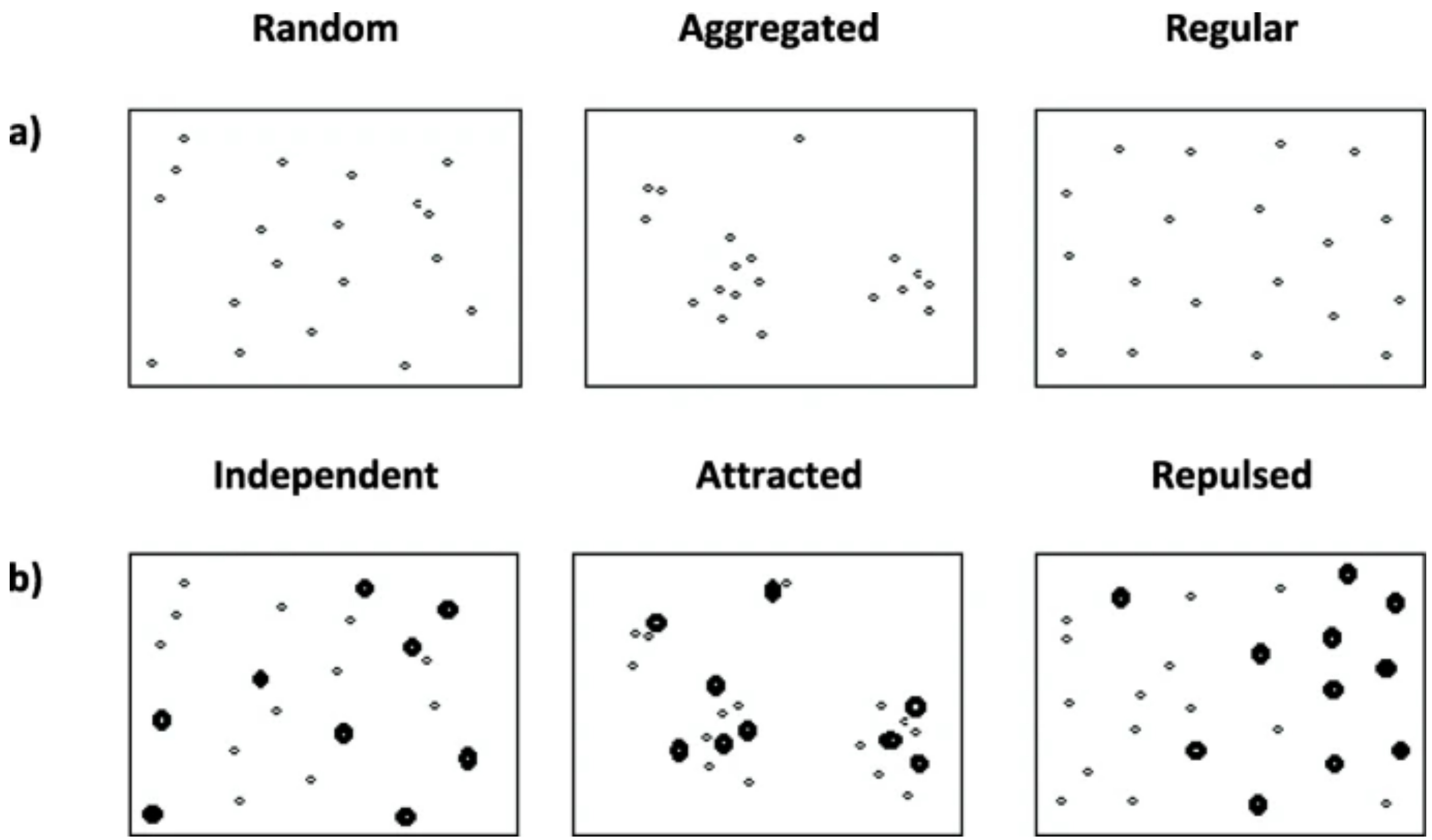
Spatial Data Mining:

Point Pattern Analysis

Trajectory Analysis

Point Pattern Analysis

Point Pattern Analysis is a data mining technique used to extract meaningful information from datasets containing spatial data points.



Common Questions:

- What does the pattern look like?
- What is the nature of the distribution of points?
- Why do events occur in those places and not in others?

Source: Geographic Data Science with Python; Ben-Said, M. (2021).

Point Pattern Analysis

Point Pattern Analysis is a data mining technique used to extract meaningful information from datasets containing spatial data points.

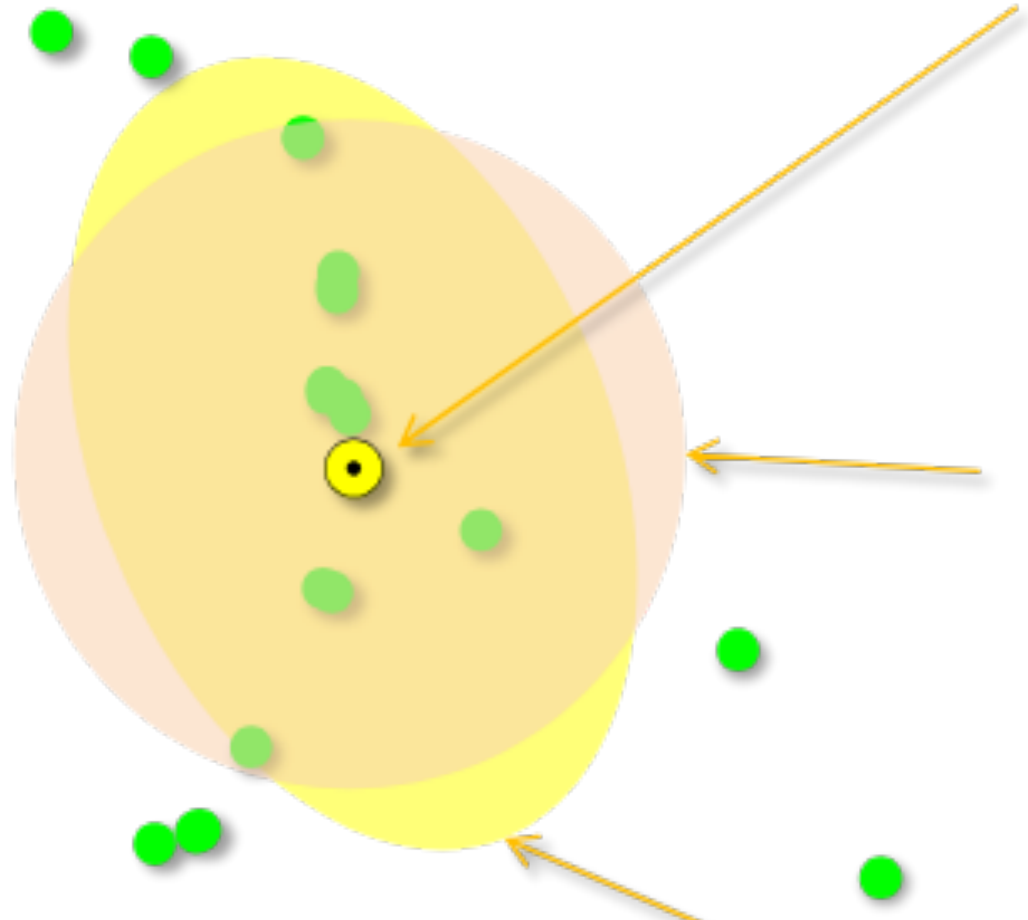
Methods:

- Centrography
  - Summary statistics on mean centre, standard distance and standard deviational ellipse
- Density-based analysis
- Distance-based analysis

Source: Point Pattern Analysis by Manuel Gimond

Point Pattern Analysis

Centrography is the analysis of centrality in a point pattern.



Mean center is the computed average X and Y coordinate values.

$$\bar{s} = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n} \right)$$

measure of tendency

Standard distance is a measure of the variance between the average distance of the features to the mean center.

$$d = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_x)^2 + (y_i - \mu_y)^2}{n}}$$

measures of dispersion

Standard deviational ellipse computes separate standard distances for each axis.

$$d_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n}}$$

$$d_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \mu_y)^2}{n}}$$

These measures provide a summary of an entire pattern, but they tell us little about the spatial organisation of each point.

Source: Point Pattern Analysis by Manuel Gimond

Point Pattern Analysis

Density-based analysis

How the points are distributed relative to the study extent – a **first-order** property of the point pattern.

- Global Density
- Local Density
  - Quadrat Density
  - Kernel Density

Distance-based analysis

How the points are distributed relative to one another - a **second-order** property of the point pattern.

- Average Nearest Neighbour
- K and L functions
- Pair Correlation Function

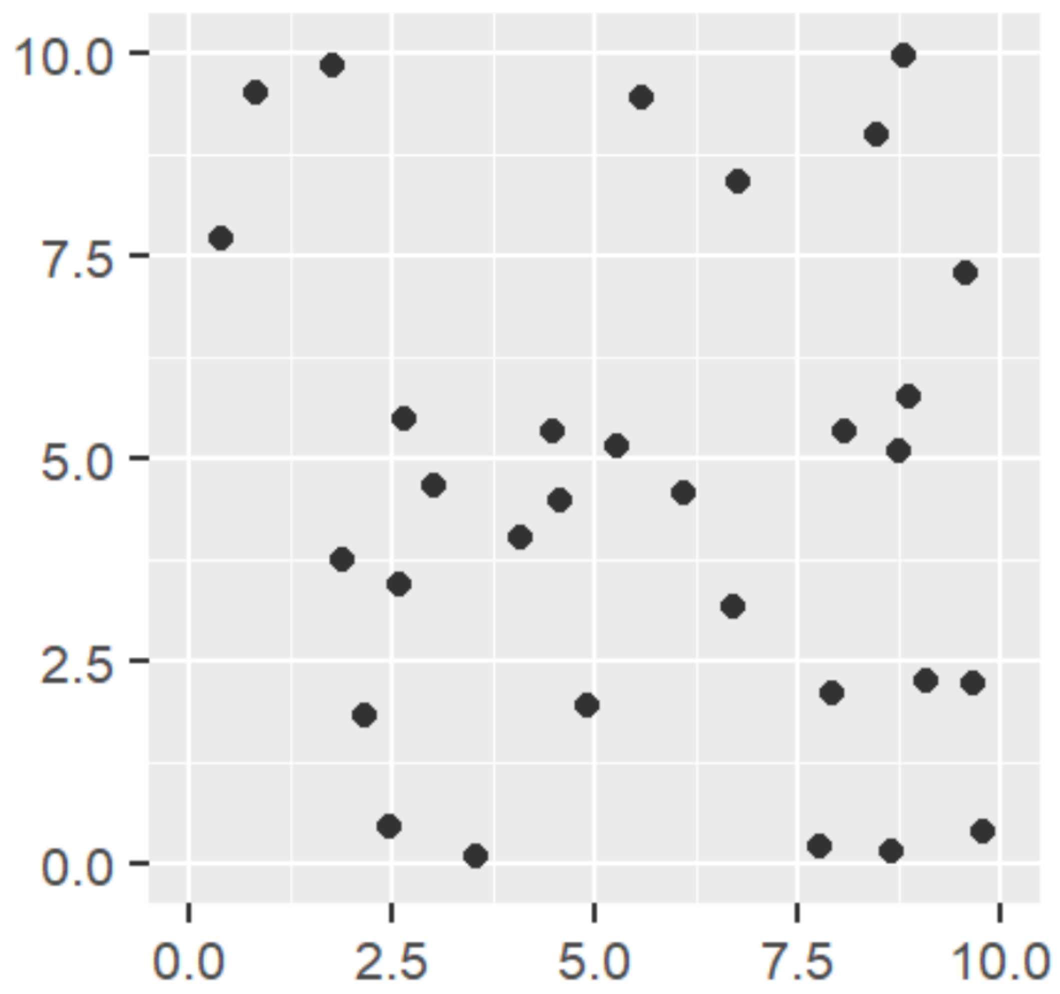
These statistical devices help us in characterising whether a point pattern is spatially clustered or dispersed.

Source: Point Pattern Analysis by Manuel Gimond

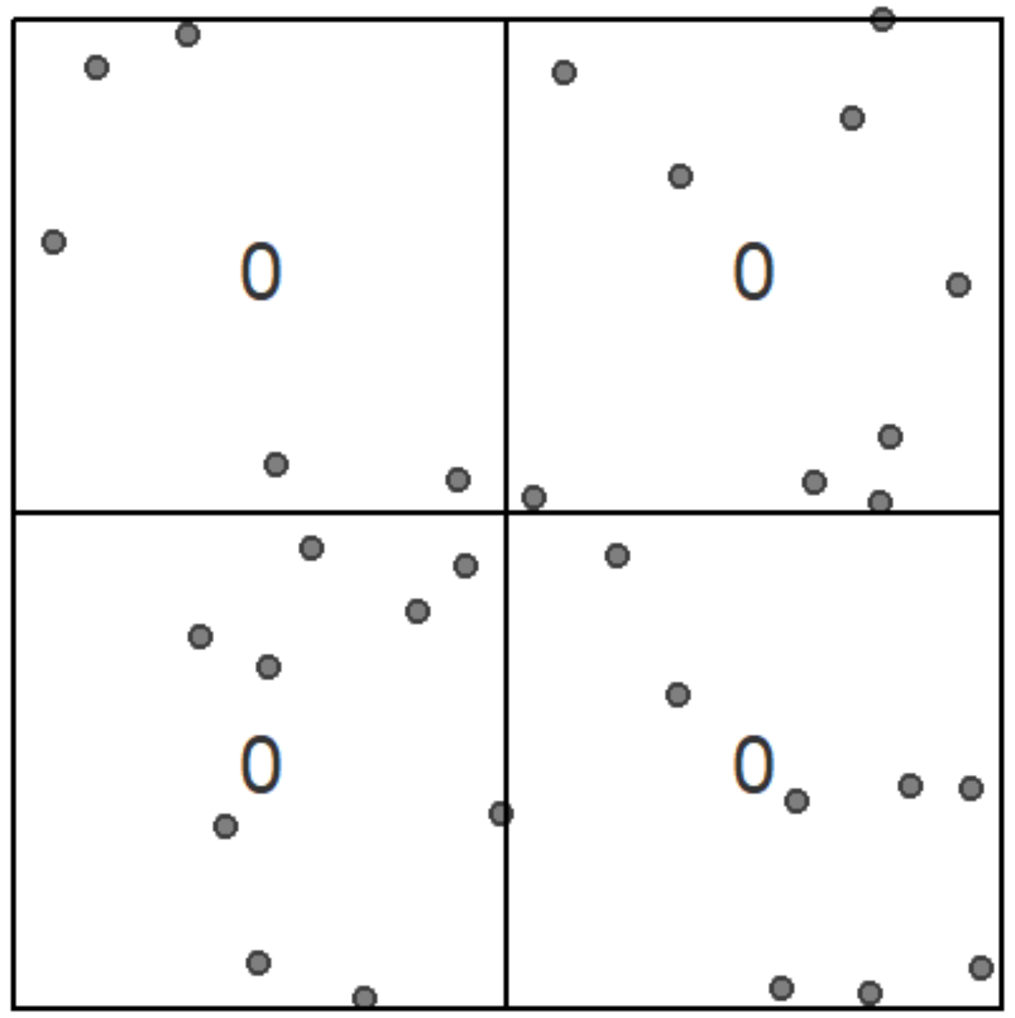
Point Pattern Analysis: Density-based

- Global Density

$\widehat{\lambda} = \frac{n}{a}$

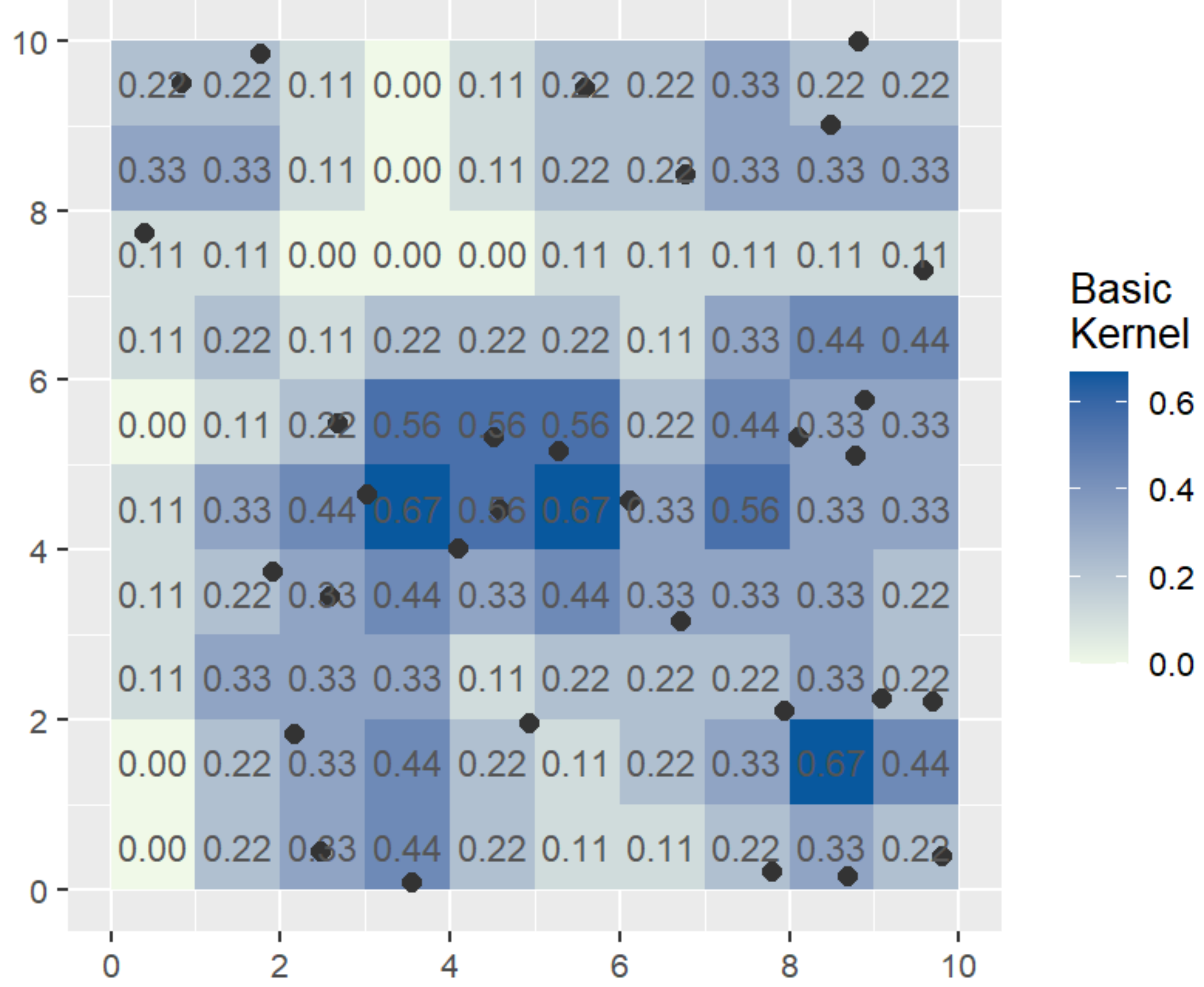


- Quadrat Density



- Kernel Density

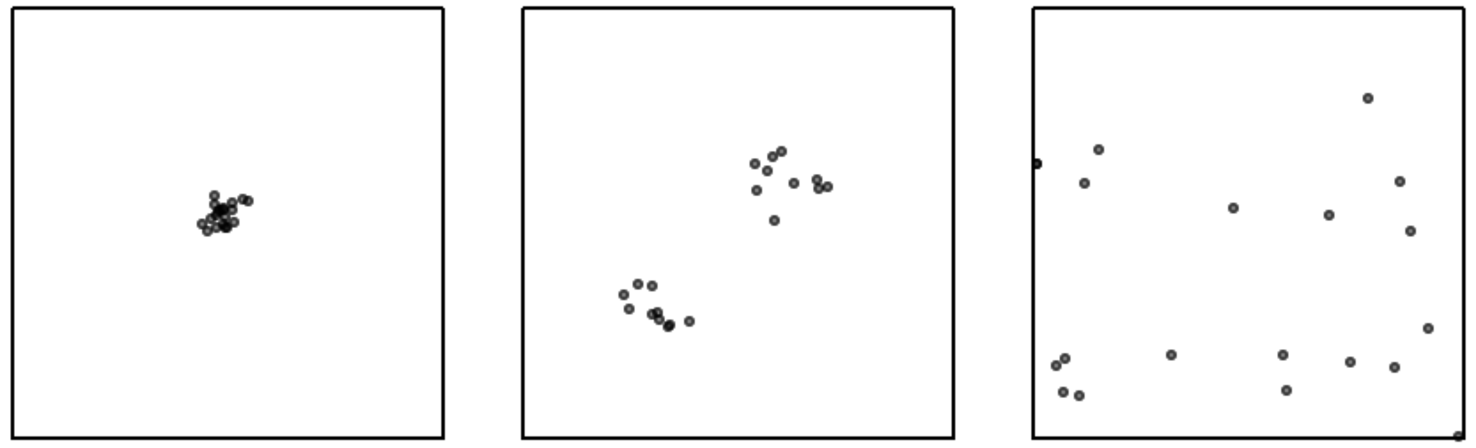
moving sub-region window

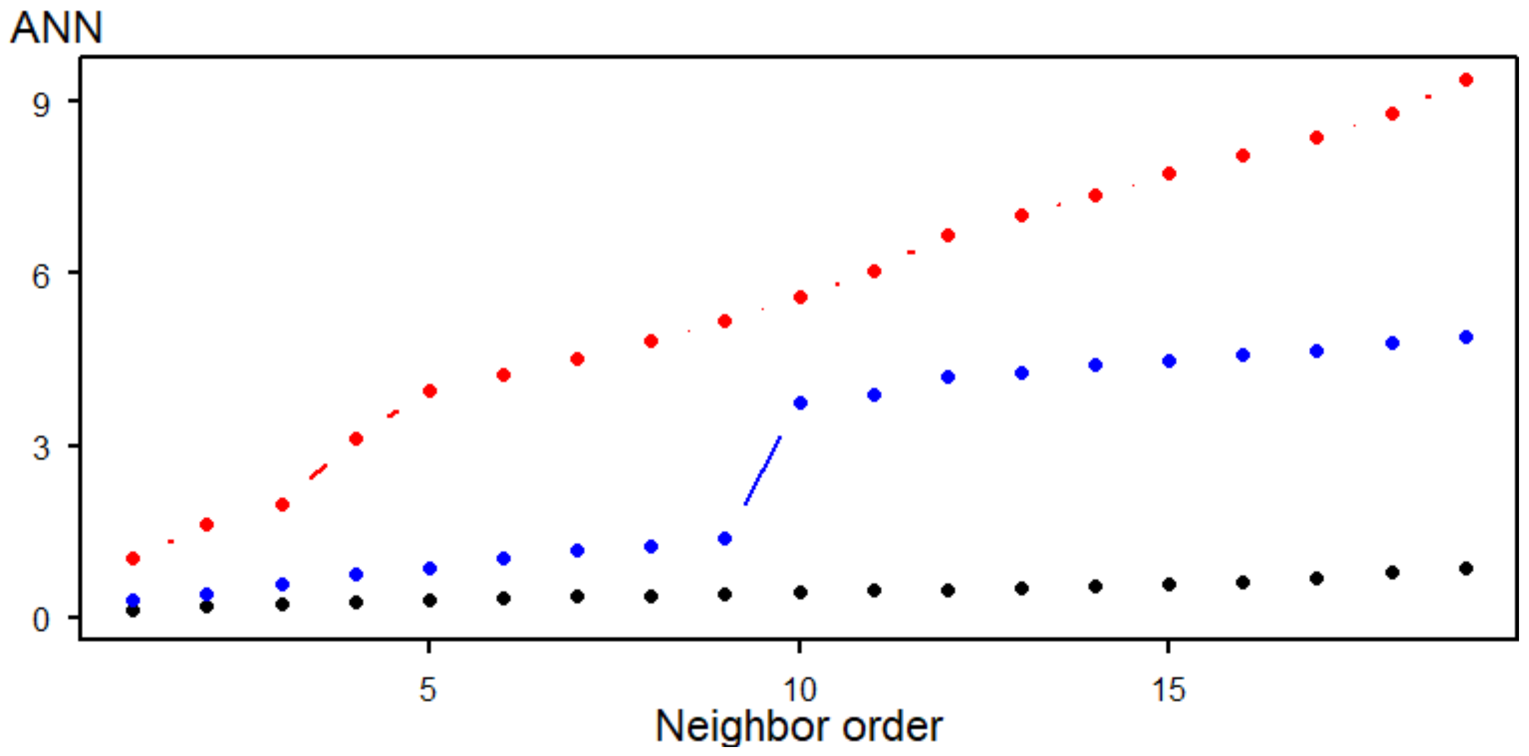


Source: Point Pattern Analysis by Manuel Gimond

Point Pattern Analysis: Distance-based

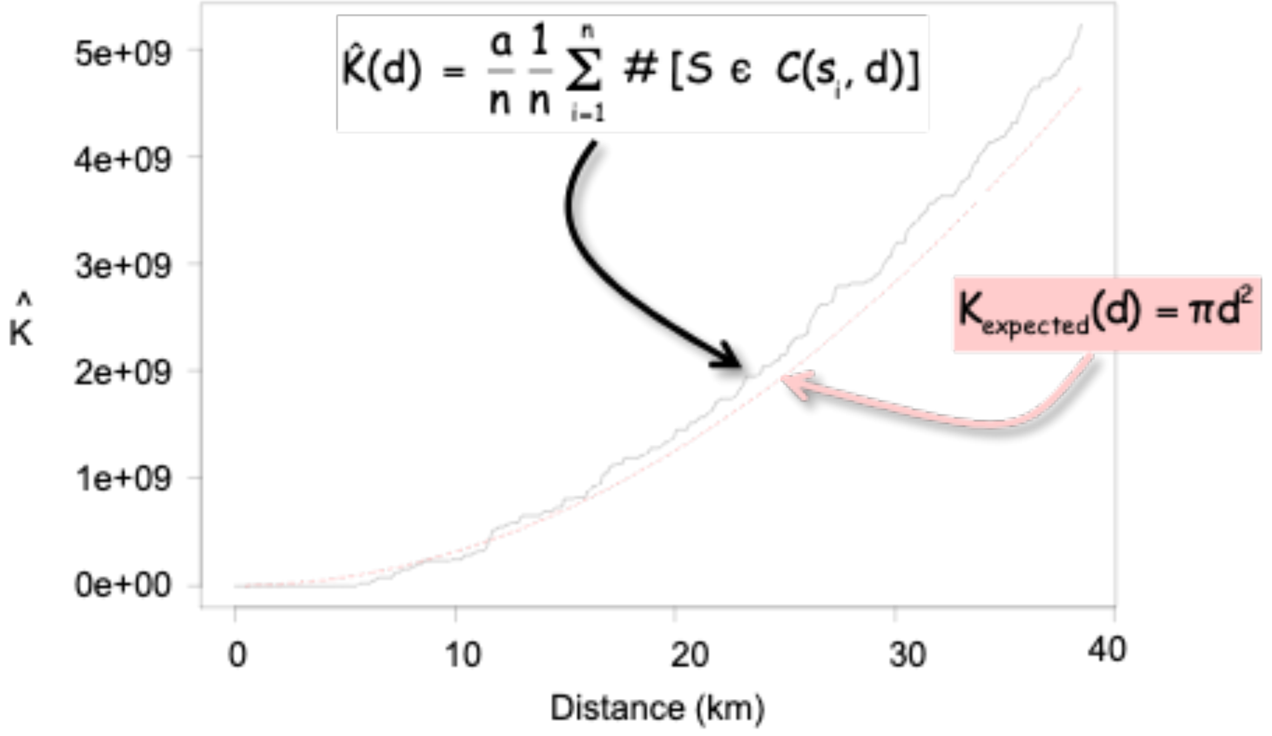
- Average Nearest Neighbour



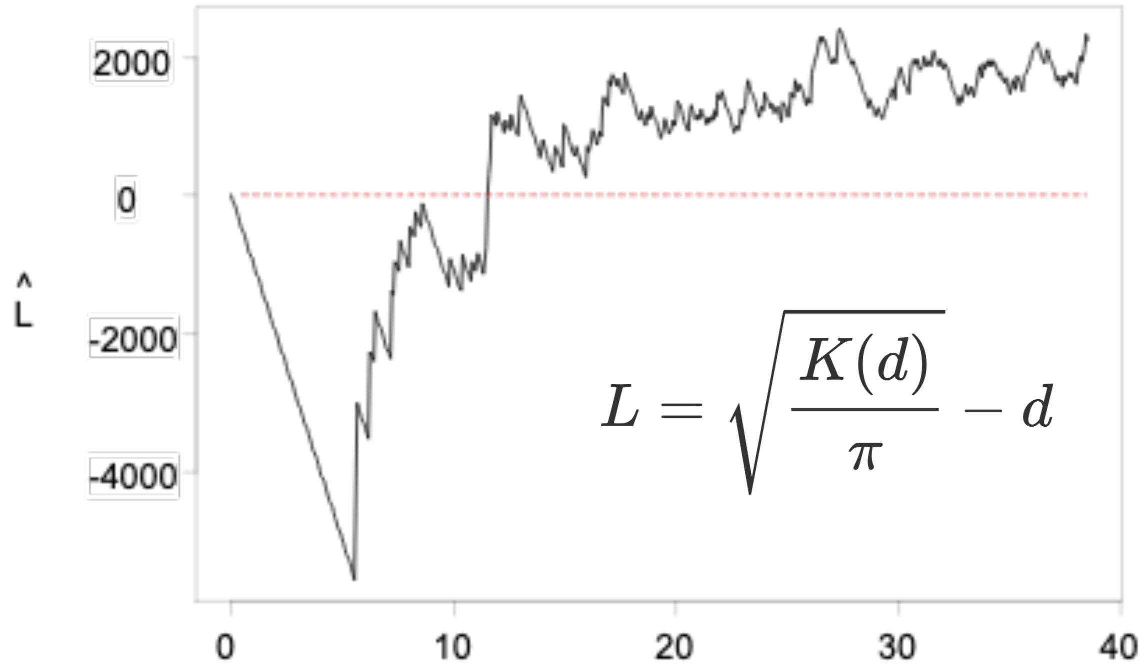


- K function

summarises the distance between points for *all* distances

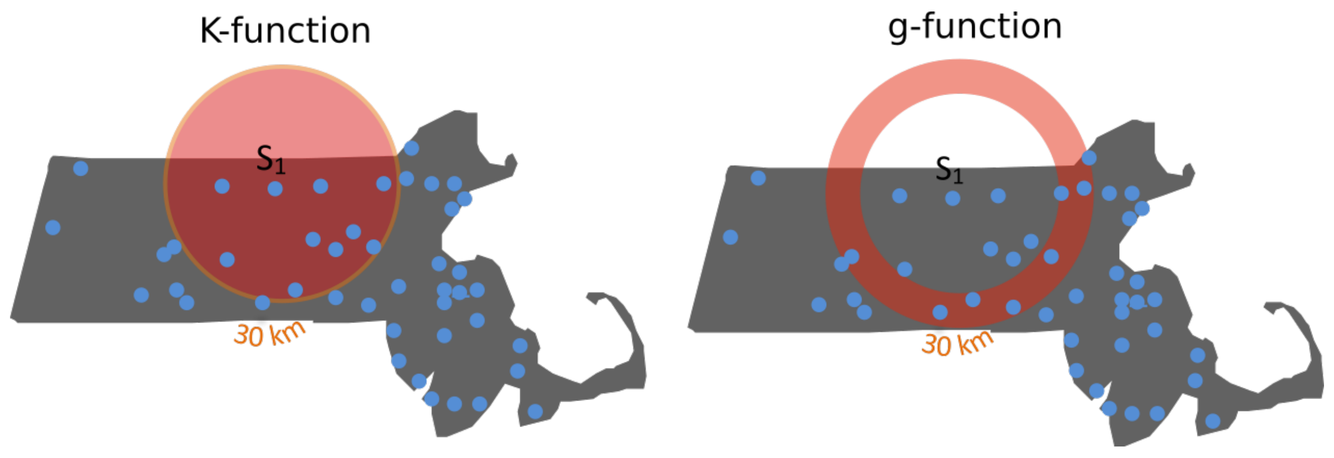


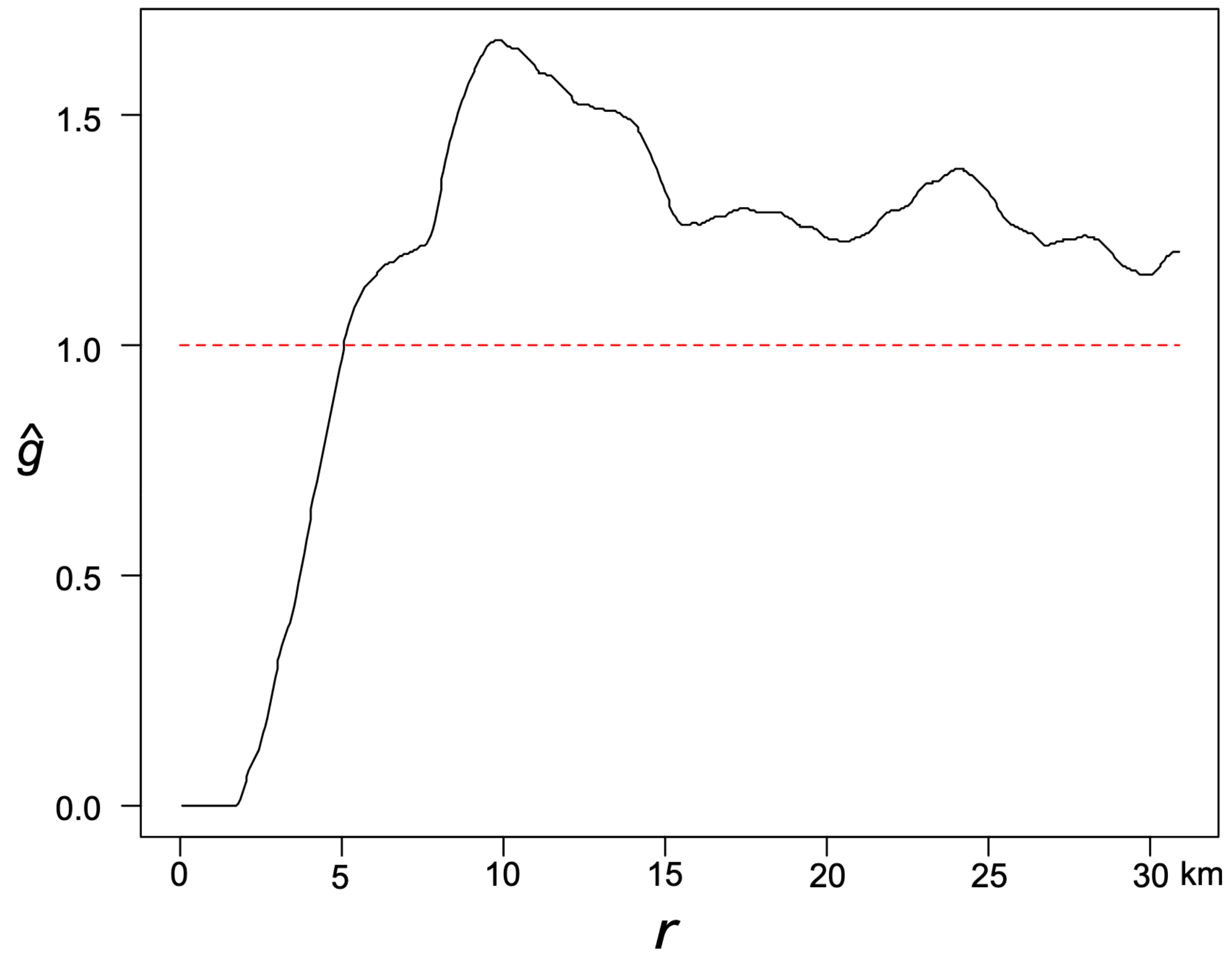
- L function



- The Pair Correlation Function g

not cumulative as K

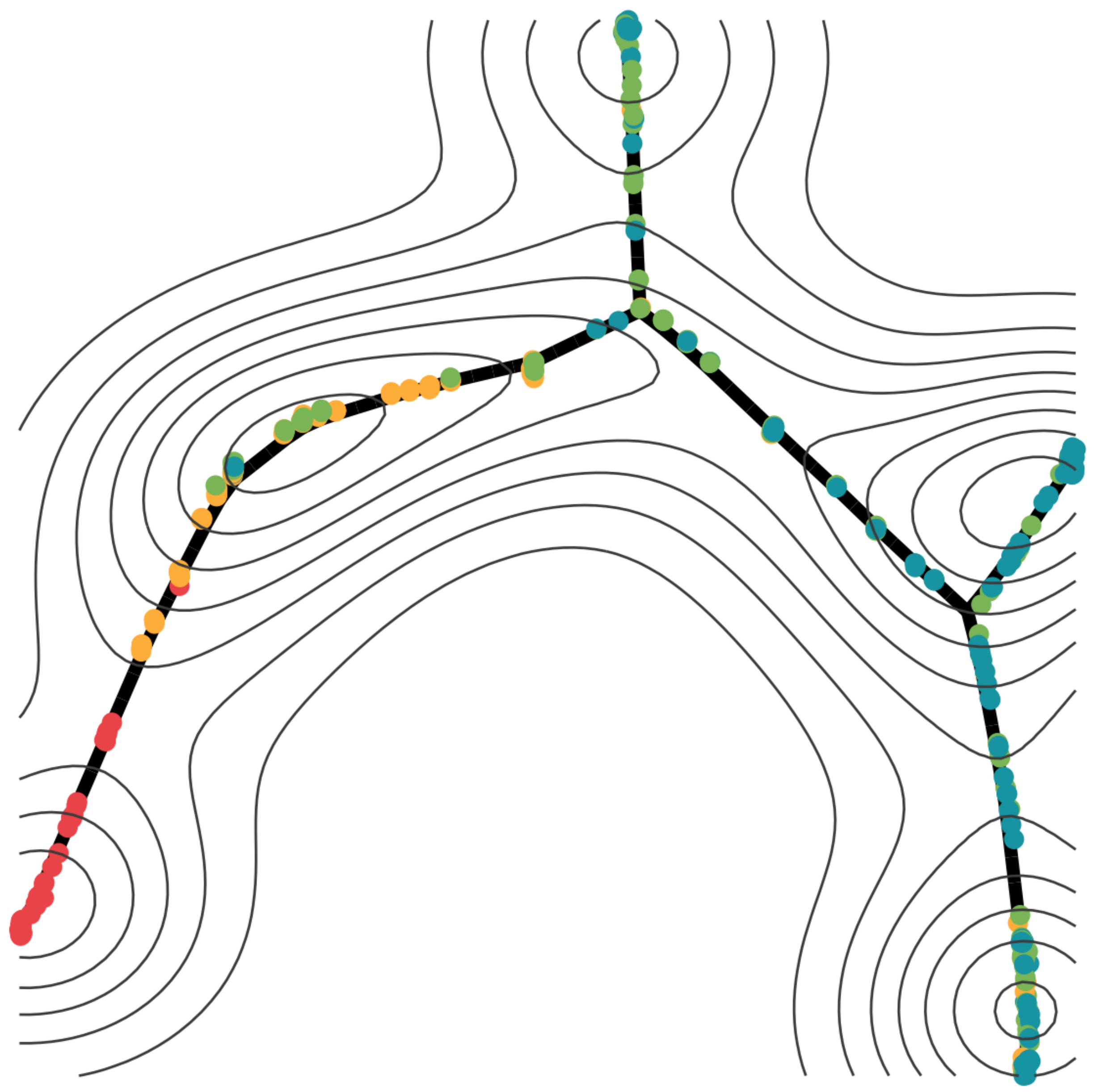




Source: Point Pattern Analysis by Manuel Gimond

Trajectory Analysis

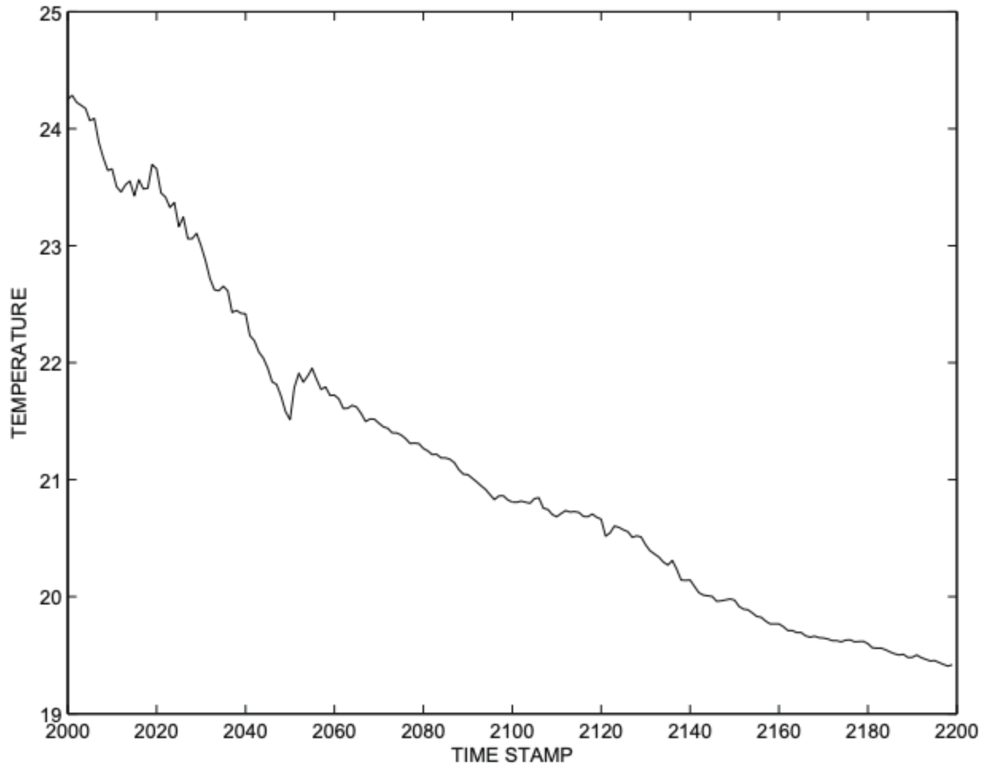
- GPS-enabled devices, such as mobile phones, has enabled the large-scale collection of trajectory data.
- Trajectory data can be analysed for a very wide variety of insights, such as determining co-location patterns, clusters and outliers.
- Trajectory data is different from the other kinds of spatial data, because its key attribute is time -> it is spatiotemporal data.



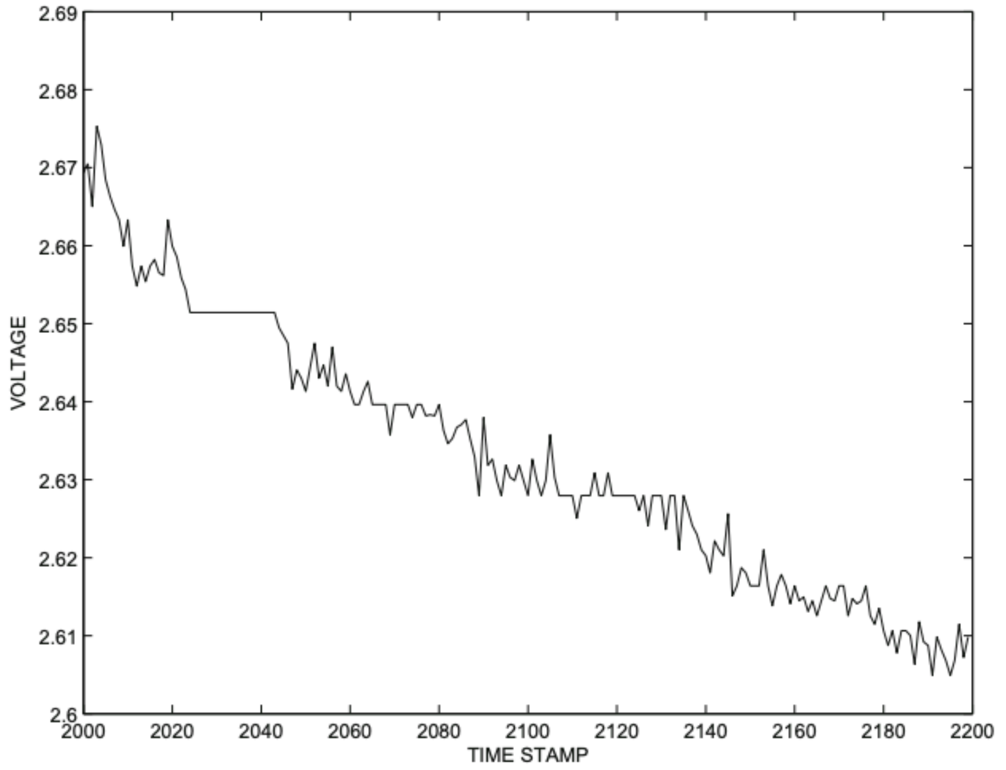
Source: Charu C. Aggarwal. Data mining: the textbook.

Trajectory Transformation

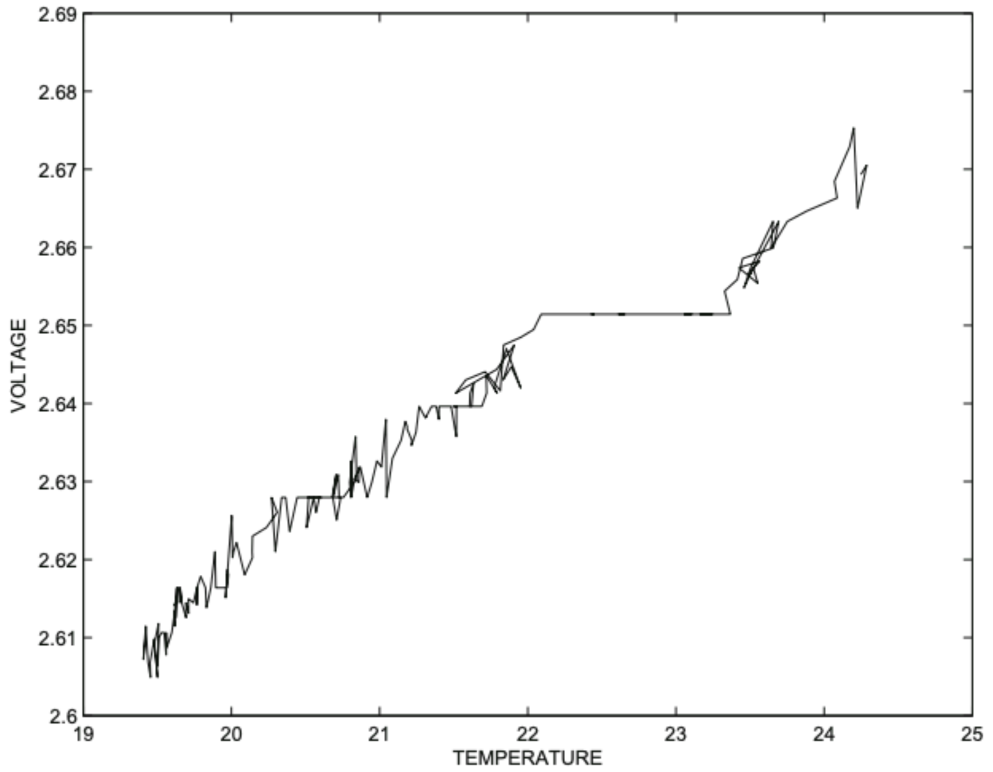
- Trajectory data is a form of multivariate time series data.
- For a trajectory in two dimensions, the X-coordinate and Y-coordinate of the trajectory form two components of the multivariate series.



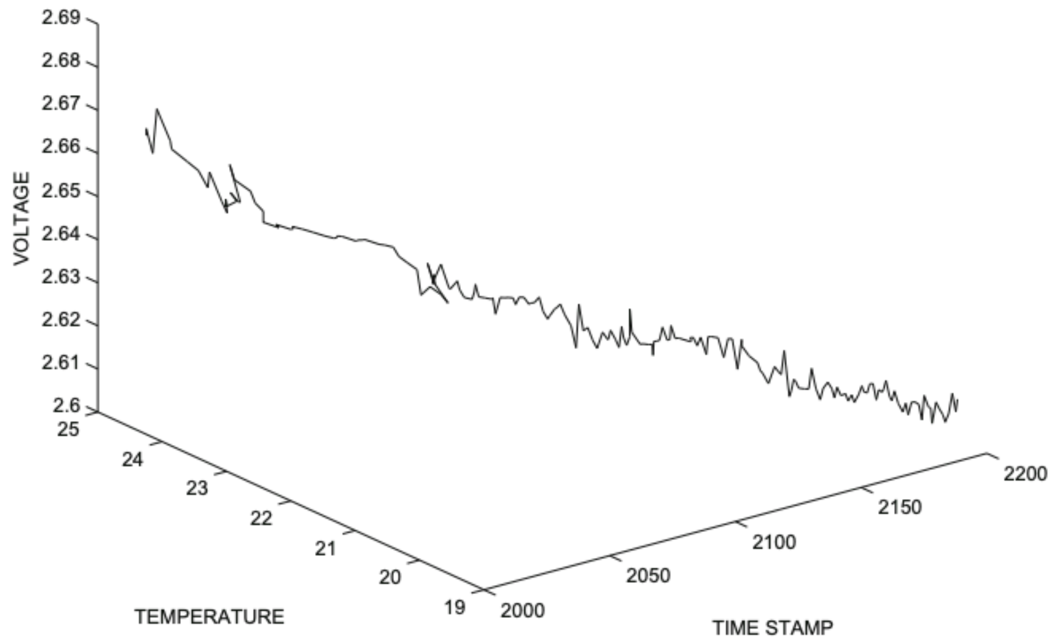
(a) Temperature



(b) Voltage



(c) Temperature-Voltage Trajectory

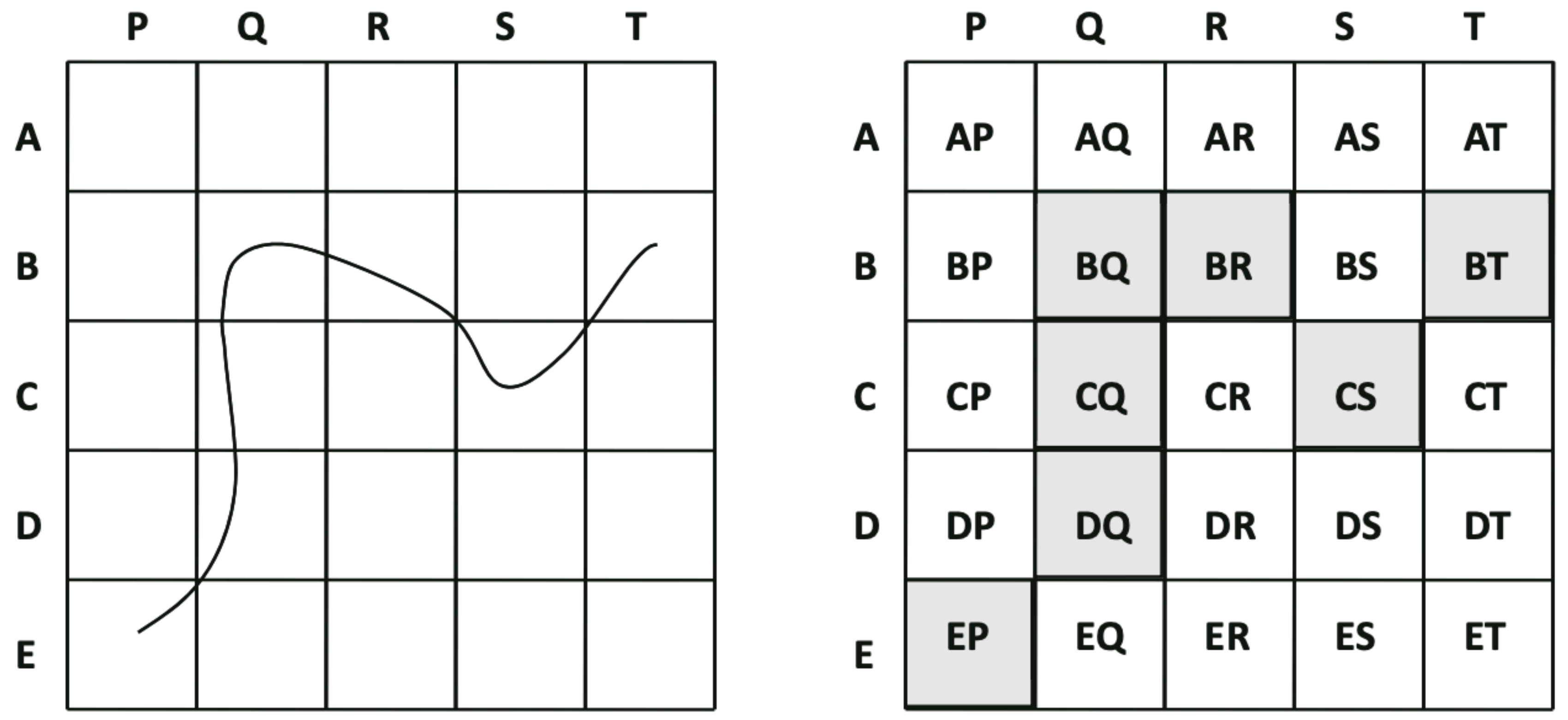


(d) Time-Temperature-Voltage Trajectory

Source: Charu C. Aggarwal. Data mining: the textbook.

Trajectory Pattern Mining

- Frequent Trajectory Paths
  - Transform the multidimensional trajectory to a 1-dimensional discrete distance - *spatial tile transformation* (via grid-based discretisation, for example)
  - Can apply any *sequential pattern mining* algorithm after
  - Can also introduce time dimension - *spatiotemporal tile transformation*



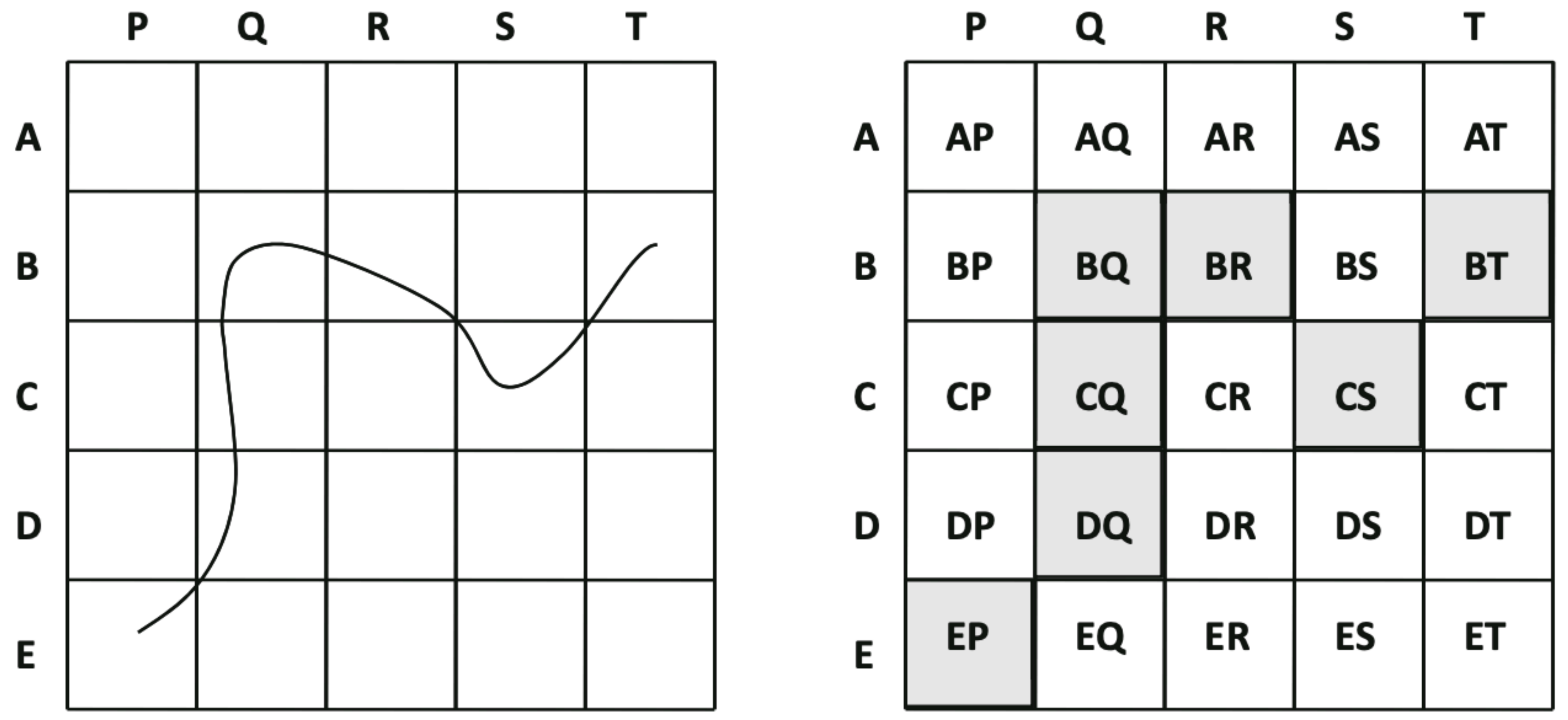
(a) Trajectory (b) Relevant grid regions

$$EP, DQ, CQ, BQ, BR, CS, BT$$

$$EP:1, EP:2, DQ:2, DQ:3, DQ:4, CQ:5, BQ:5, BR:5, CS:6, CS:7, BT:7$$

Trajectory Pattern Mining

- Colocation Patterns
  - Designed to discover *social connections* between the *trajectories of different individuals*: individuals who frequently appear at the same point at the same time are likely to be related to one another
  - Can apply any *frequent pattern mining* algorithm after



(a) Trajectory (b) Relevant grid regions

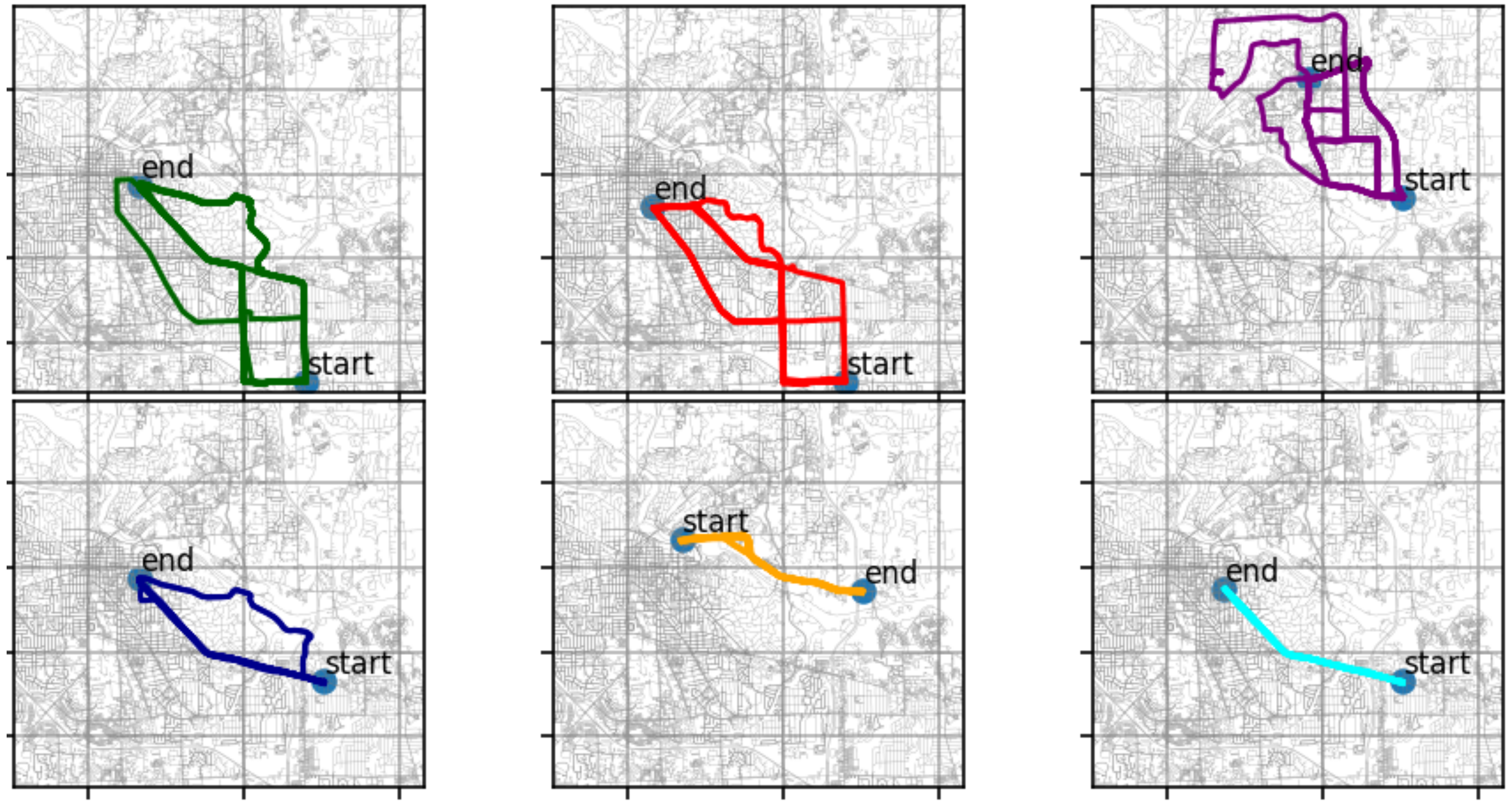
$$EP, DQ, CQ, BQ, BR, CS, BT$$

$$EP:1, EP:2, DQ:2, DQ:3, DQ:4, CQ:5, BQ:5, BR:5, CS:6, CS:7, BT:7$$

$$EP:5 \Rightarrow \{3, 9, 11\}$$

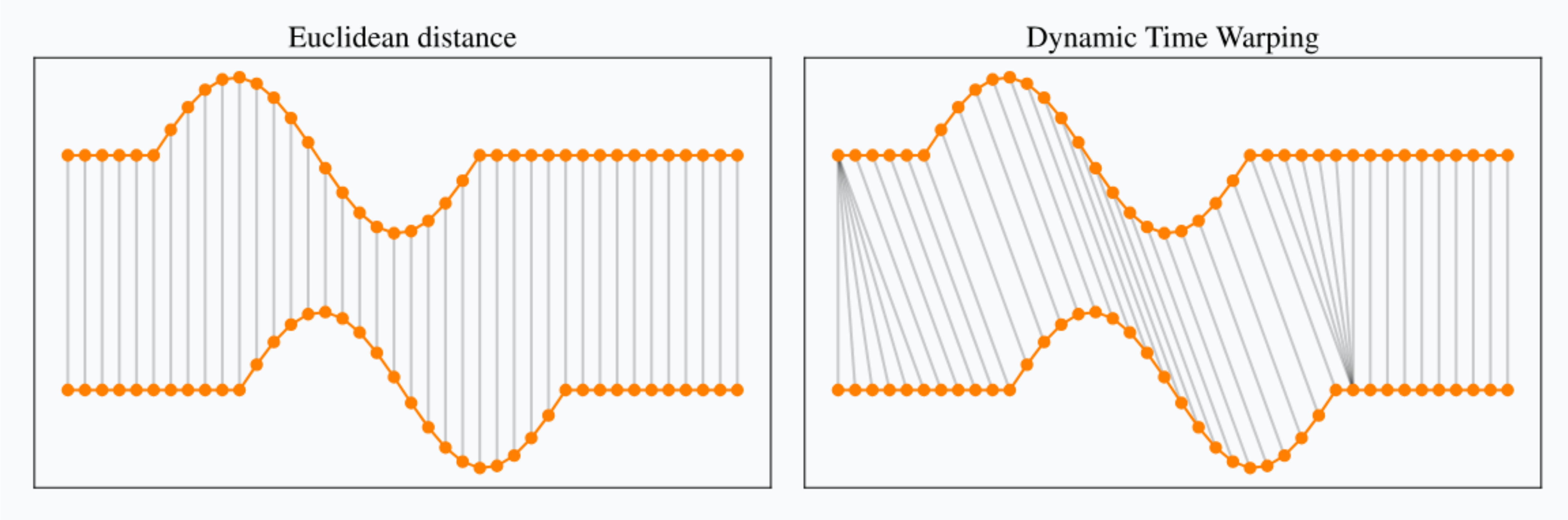
Trajectory Clustering

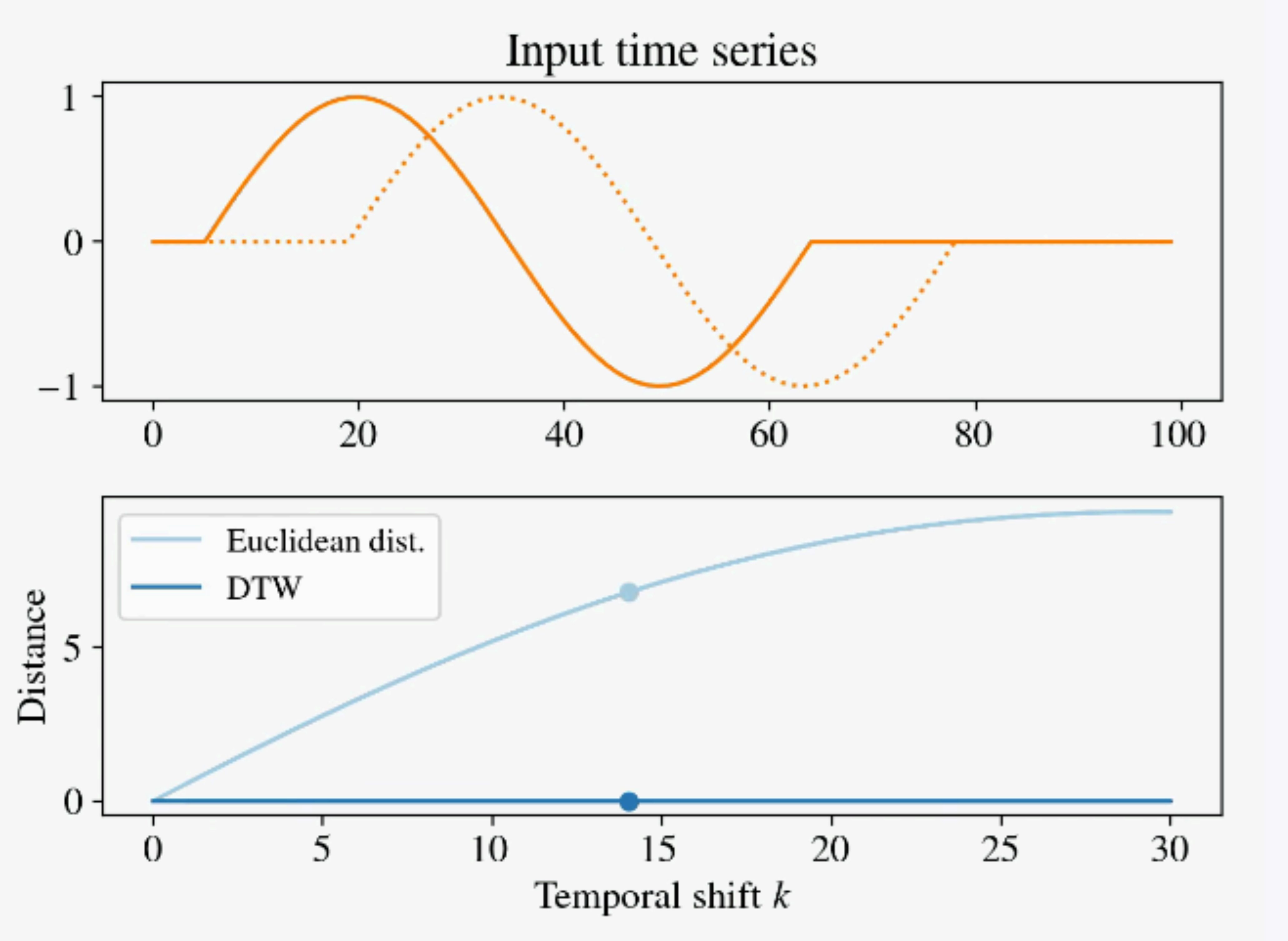
- Conventional clustering algorithms, with the use of distance function between trajectories.
  - Once a distance function is defined, can apply k-medoids, graph-based methods, or others.
- Converting trajectories into sequences of discrete symbols.
  - Segment extraction, grid-based discretisation, etc.
  - After the transformation, pattern mining algorithms are applied to the extracted sequence of symbols.



Trajectory Clustering: Computing Similarity

- Similarity computation between trajectories is not very different from that of time series data.
- DTW - Dynamic Time Wrapping - seeks for the temporal alignment (matching between time indexes of the two time series) that minimises Euclidean distance between aligned series





Sources: Charu C. Aggarwal. Data mining: the textbook;
An introduction to DTW by Romain Tavenard

Trajectory Clustering: Computing Similarity

- Similarity computation between trajectories is not very different from that of time series data.
- DTW - Dynamic Time Warping
- MDTW - multidimensional DTW - the only difference from the case of univariate time series data is the substitution of the 1-dimensional distances in the recursion with 2-dimensional distances.

$$DTW(i,j) = distance(x_i, y_j) + \min \begin{cases} DTW(i, j-1) & \text{repeat } x_i \\ DTW(i-1, j) & \text{repeat } y_j \\ DTW(i-1, j-1) & \text{otherwise} \end{cases}$$

$$MDTW(i,j) = distance(\overline{X_i}, \overline{Y_j}) + \min \begin{cases} MDTW(i, j-1) & \text{repeat } \overline{X_i} \\ MDTW(i-1, j) & \text{repeat } \overline{Y_j} \\ MDTW(i-1, j-1) & \text{otherwise.} \end{cases}$$

Sources: Charu C. Aggarwal. Data mining: the textbook;
How DTW (Dynamic Time Warping) algorithm works | Youtube

Trajectory Clustering: Clustering Methods

- Once we have a **similarity function**, can use any method directly for any data type (k-medoids, graph-based methods).

- Alternatively, if we opt to work with a sequence, then:
  - Use grid-based discretisation to convert the N trajectories to N discrete sequences (as shown in previous slides).
  - Apply any of the sequence clustering methods to create clusters from the sequences.
  - Map the sequence clusters back to trajectory clusters.

- One advantage of the sequence clustering approach over similarity-based methods, is that many of the sequence clustering methods can ignore the irrelevant parts of the sequences in the clustering process.

Source: Charu C. Aggarwal. Data mining: the textbook.

Notebook 3

Point Pattern Analysis

Trajectory Mining

Trajectory Clustering

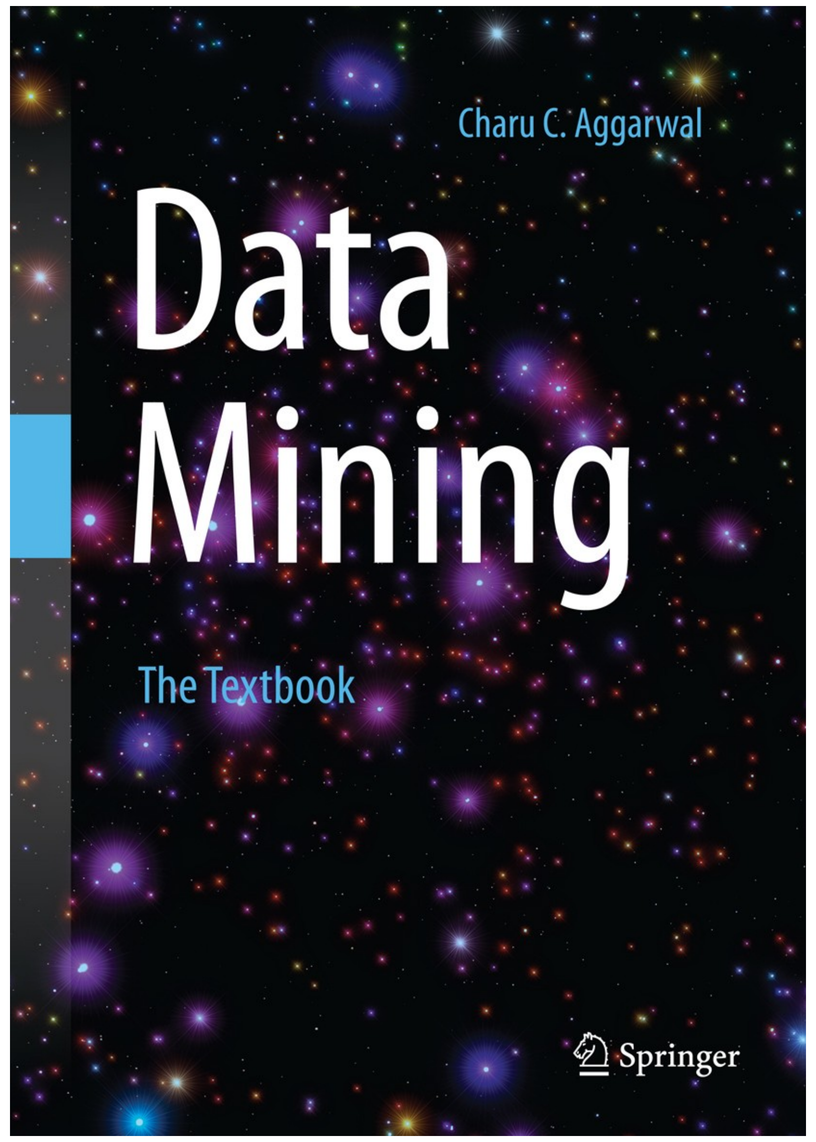
Summary

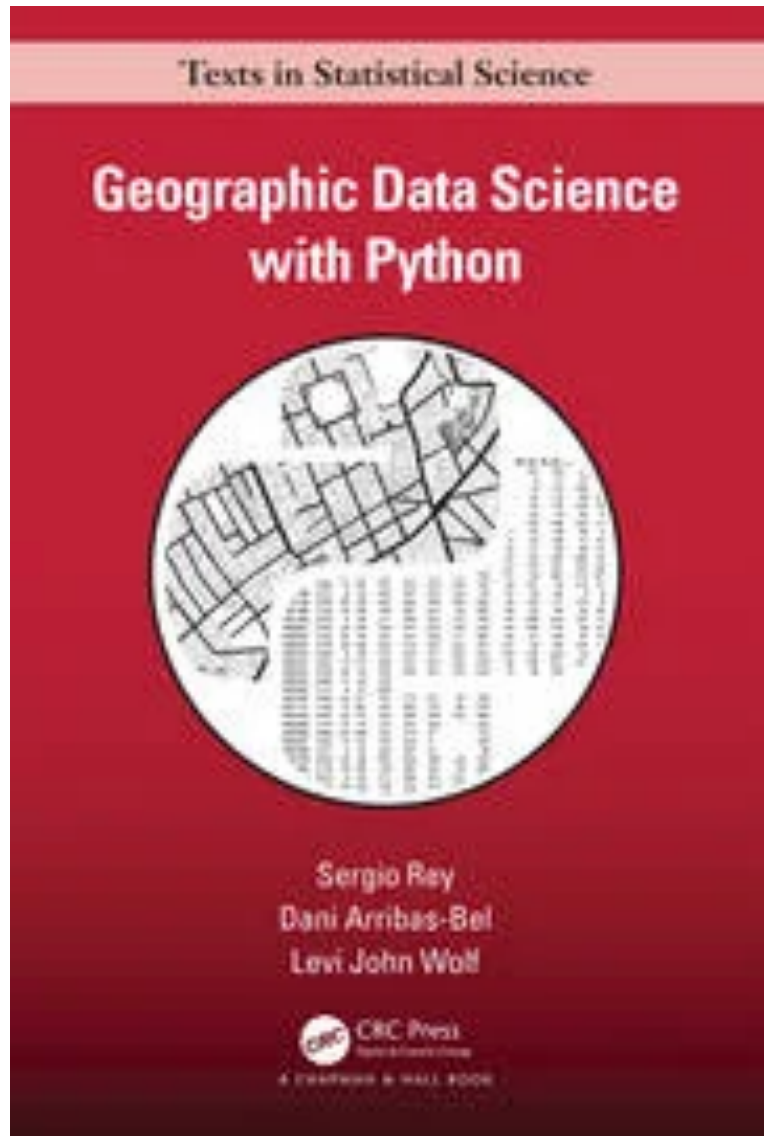
In this lecture, the goal was to familiarise ourselves with the following concepts of Spatial Data Mining: **spatial data**, **spatial autocorrelation**, **spatial clustering**, **point pattern analysis**, **trajectory analysis**.

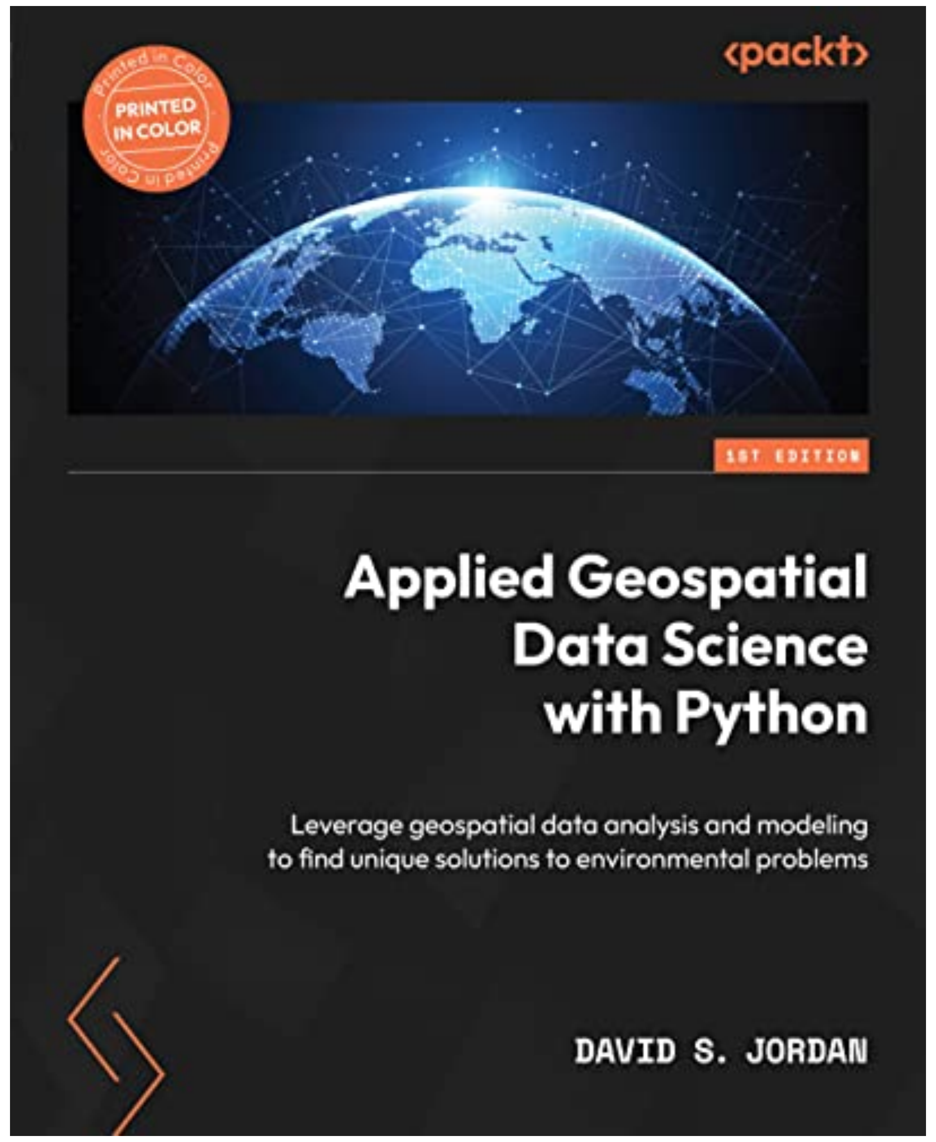
We also went beyond theoretical understanding and practiced the application of these concepts in **hand-on exercises in notebooks.**

The knowledge and skills acquired in this lecture have **broad-ranging applications**, from urban planning and environmental management to public health and transportation logistics.

Resources



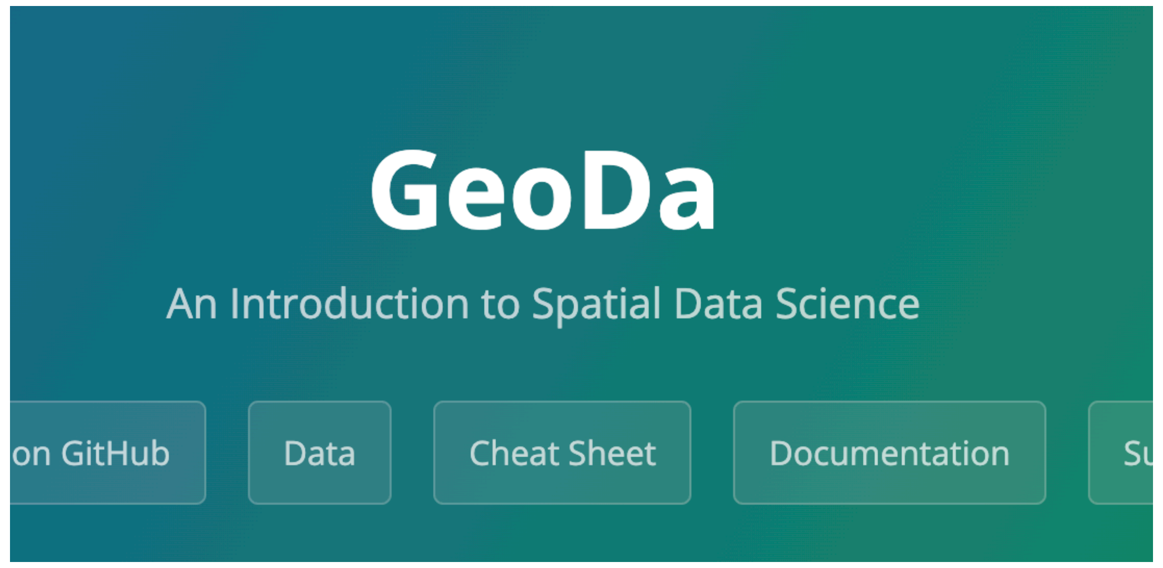


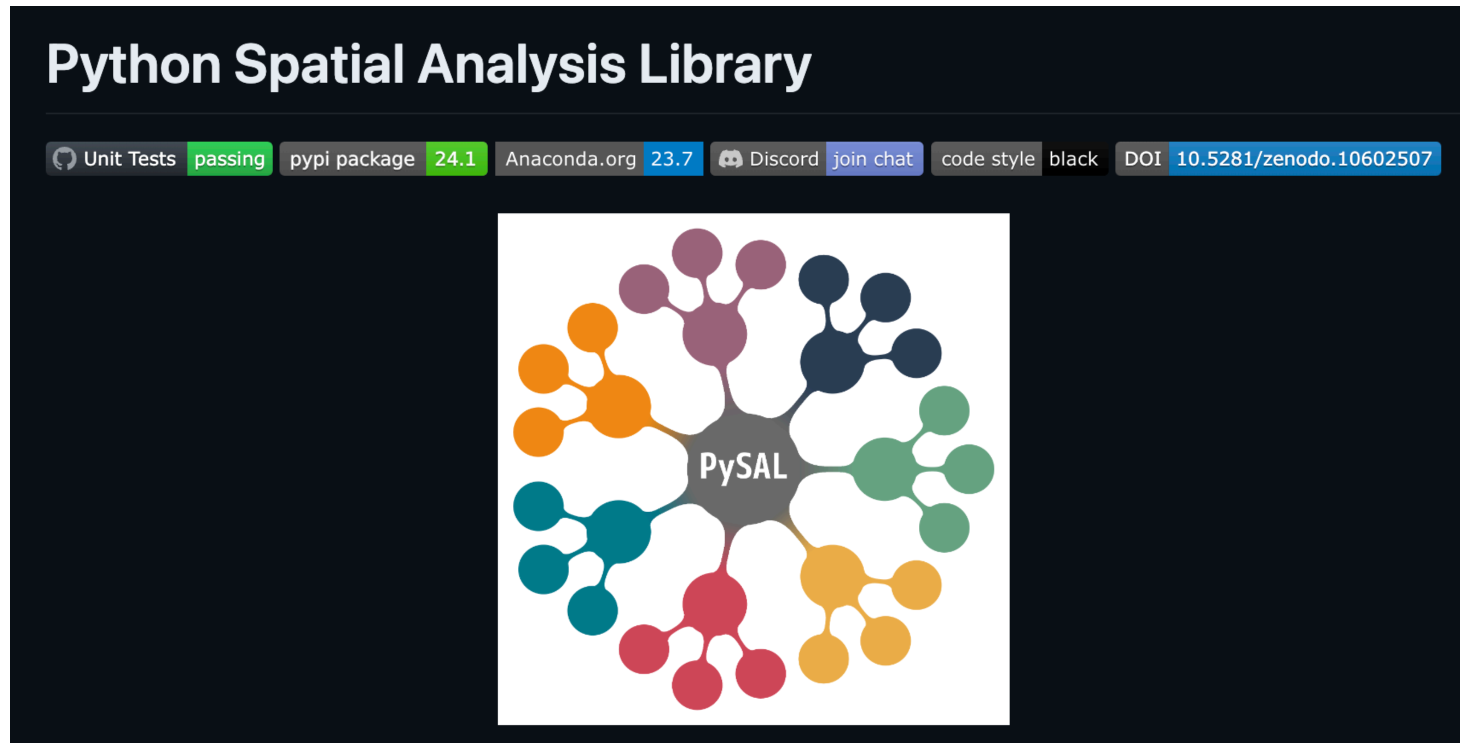


Intro to GIS and Spatial Analysis

Manuel Gimond

Last edited on 2023-12-15





Questions?

@martonkarsai

karsaim@ceu.edu naushirvanov_timur@phd.ceu.edu

https://www.martonkarsai.com/