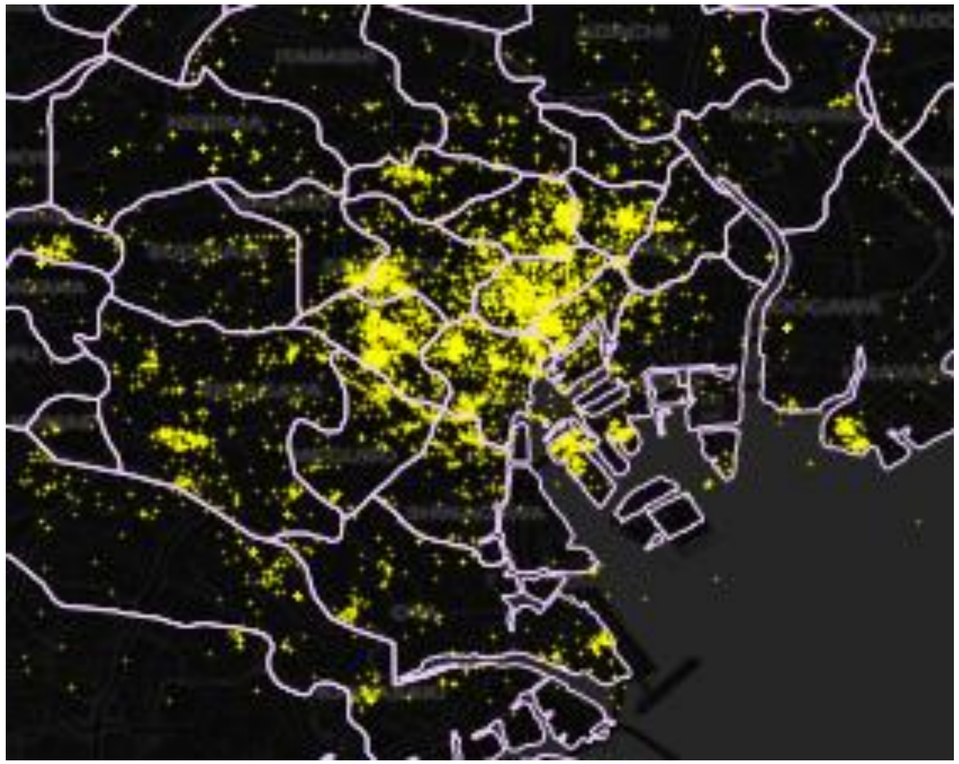
Lecture 7: Point pattern analysis

Marina Georgati

Mar 13, 2023



Objectives & agenda

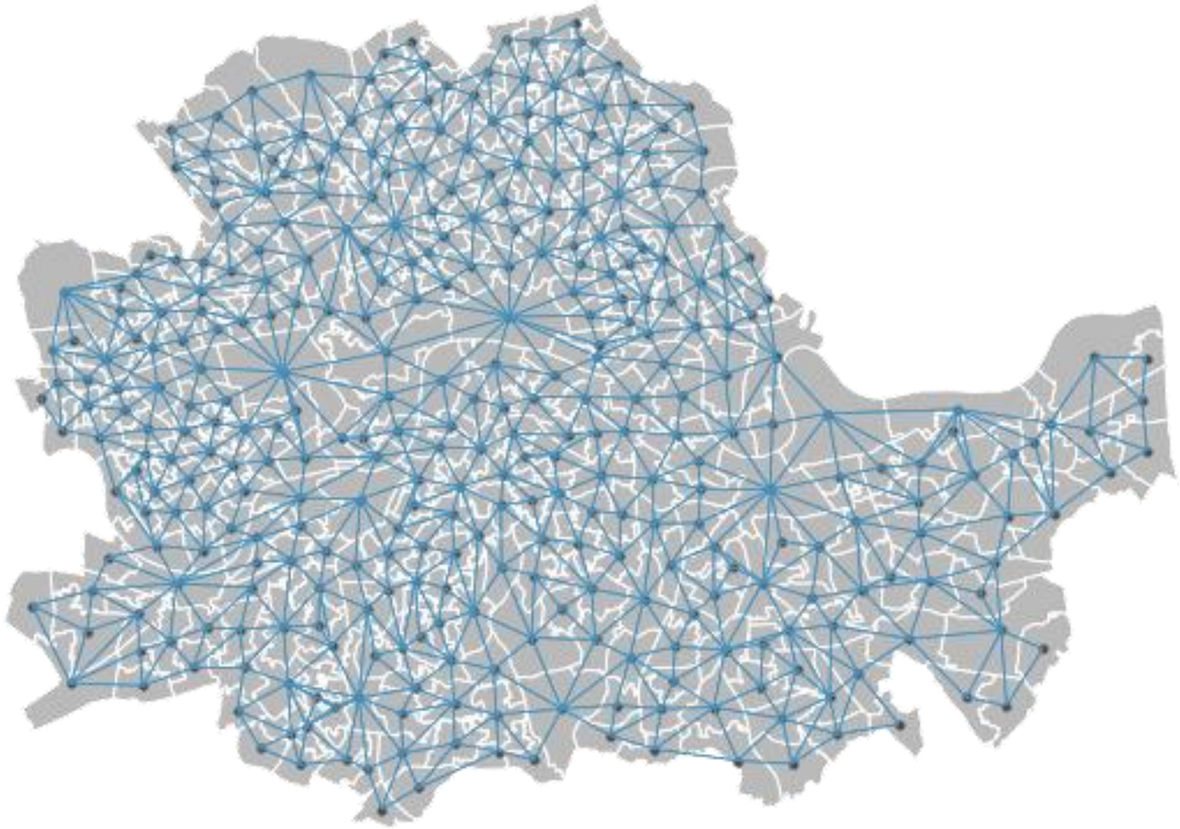
- Introduction to Point Patterns and Point Pattern Analysis
- Application examples
- Methods
  - Summary statistics
  - Density-based approaches
  - Distance-based approaches
- Hypothesis testing
- DBSCAN

Points as static objects

When location of points is fixed:

Do same analysis as with polygons

Static objects with fixed location:
cities, regions, buildings

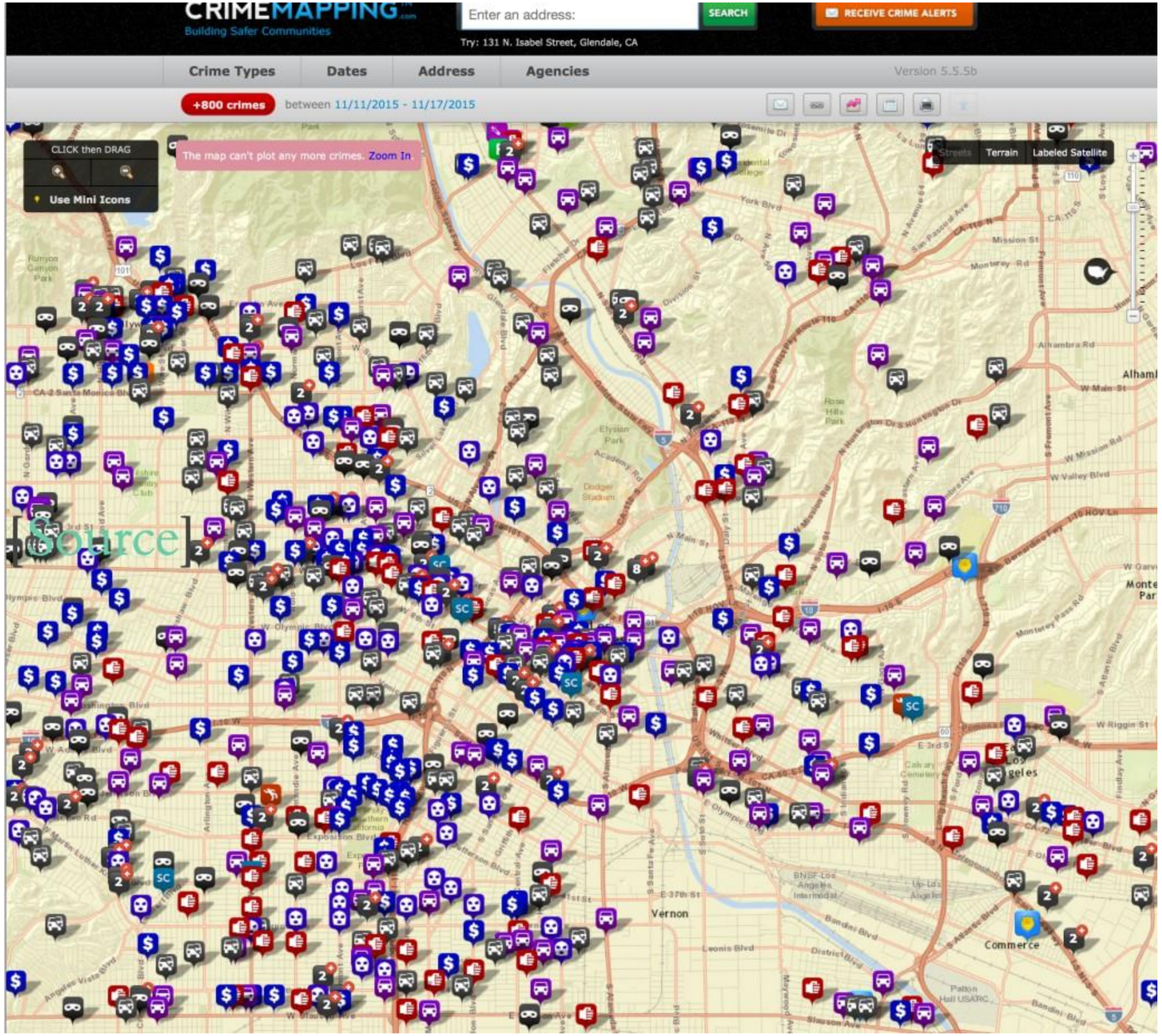


Points as events

When points stand for events, they could happen anywhere

- We try to understand the locations
- We want to characterize the spatial pattern of the points

Crime, trees, taxi pickups, tweets



What is a point pattern (PP)?

- PPs are a set of events in a study region that is represented as the simplest form of a spatial data, points 📎.

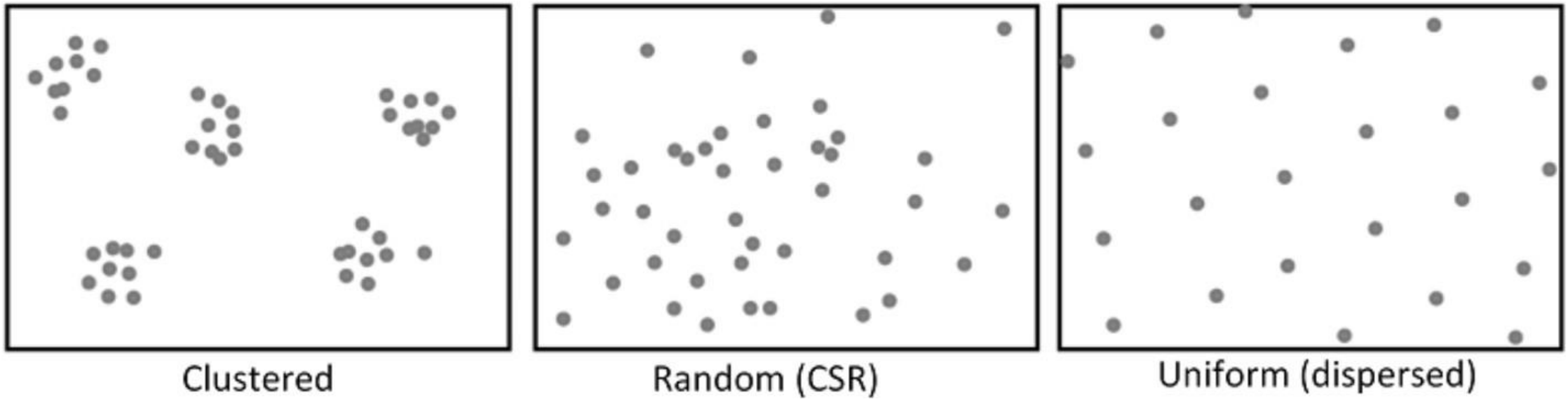
- PPA studies the spatial arrangement or distribution of those points in the study area.

Examples: Locations of trees or plants, birds / animal sightings, 911 calls, cases of a disease, wildfires, accidents, …

Point patterns

A point pattern is a distribution of points over space

These points could happen anywhere, but are only observed in certain positions



🔗 *https://gistbok.ucgis.org/bok-topics/point-pattern-analysis*

More formally…

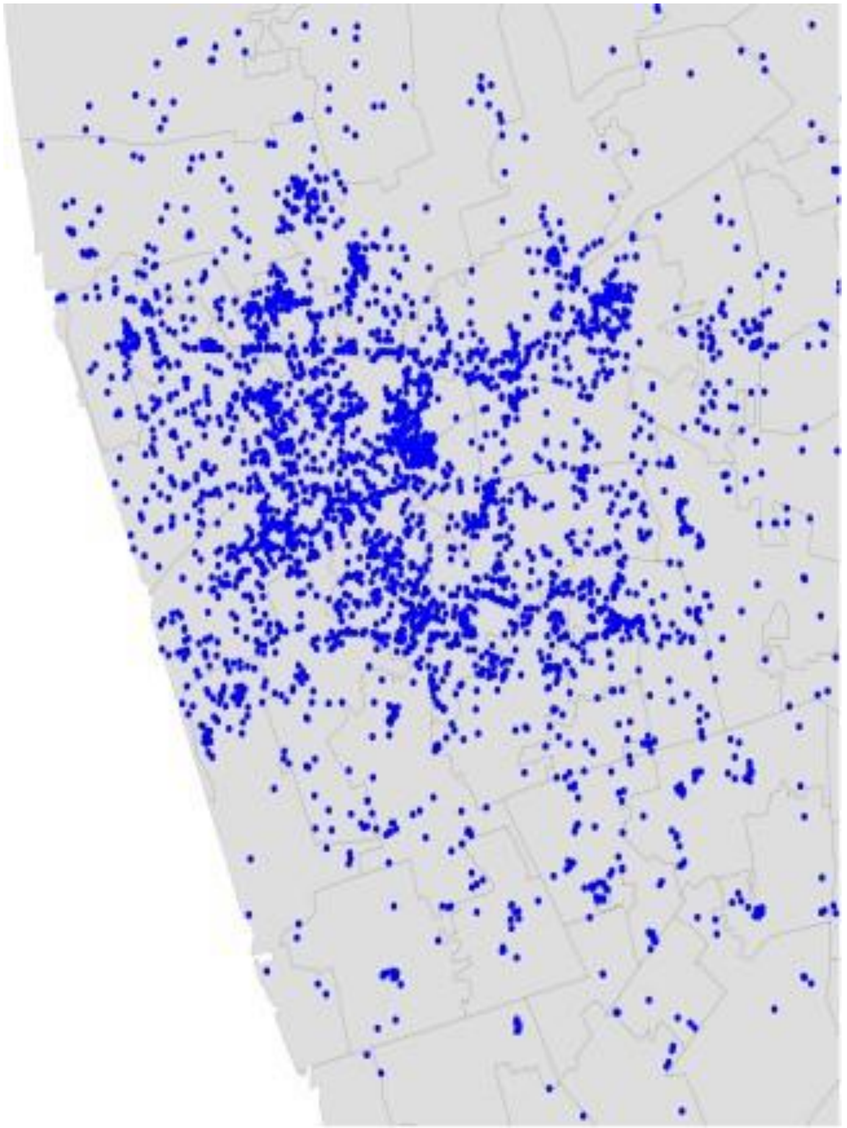
- A point pattern consists of events in a set of locations:

$$S = \{s_1, s_2, \ldots, s_i, \ldots, s_n\}$$

- Each event s_i has a location (x_i, y_i)
- Point patterns occur in study region ***A*** of area ***a***

Visualization/ analysis: One-to-one (scatter)

- Intuitive
- Effective for small data
- No aggregation = no MAUP
- Useless for large or very clustered data



Point pattern analysis (PPA)

PPA aims to describe and explain the data generating process through:

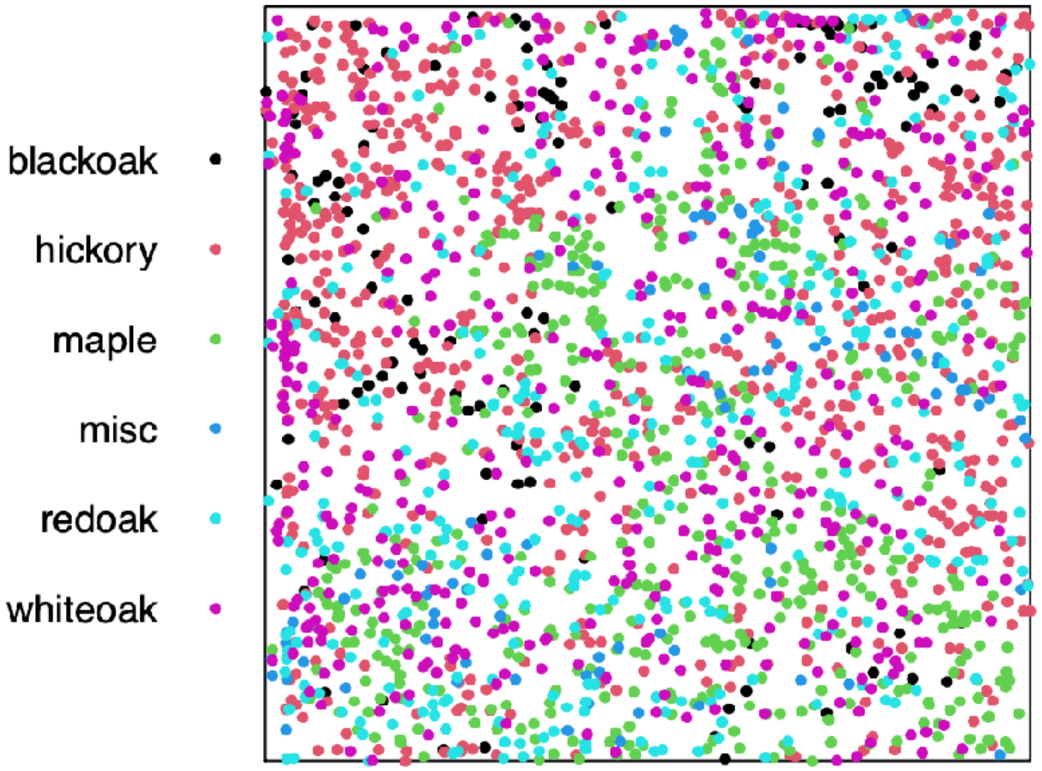
- Visualization
- Modeling the underlying process
- Clustering algorithms

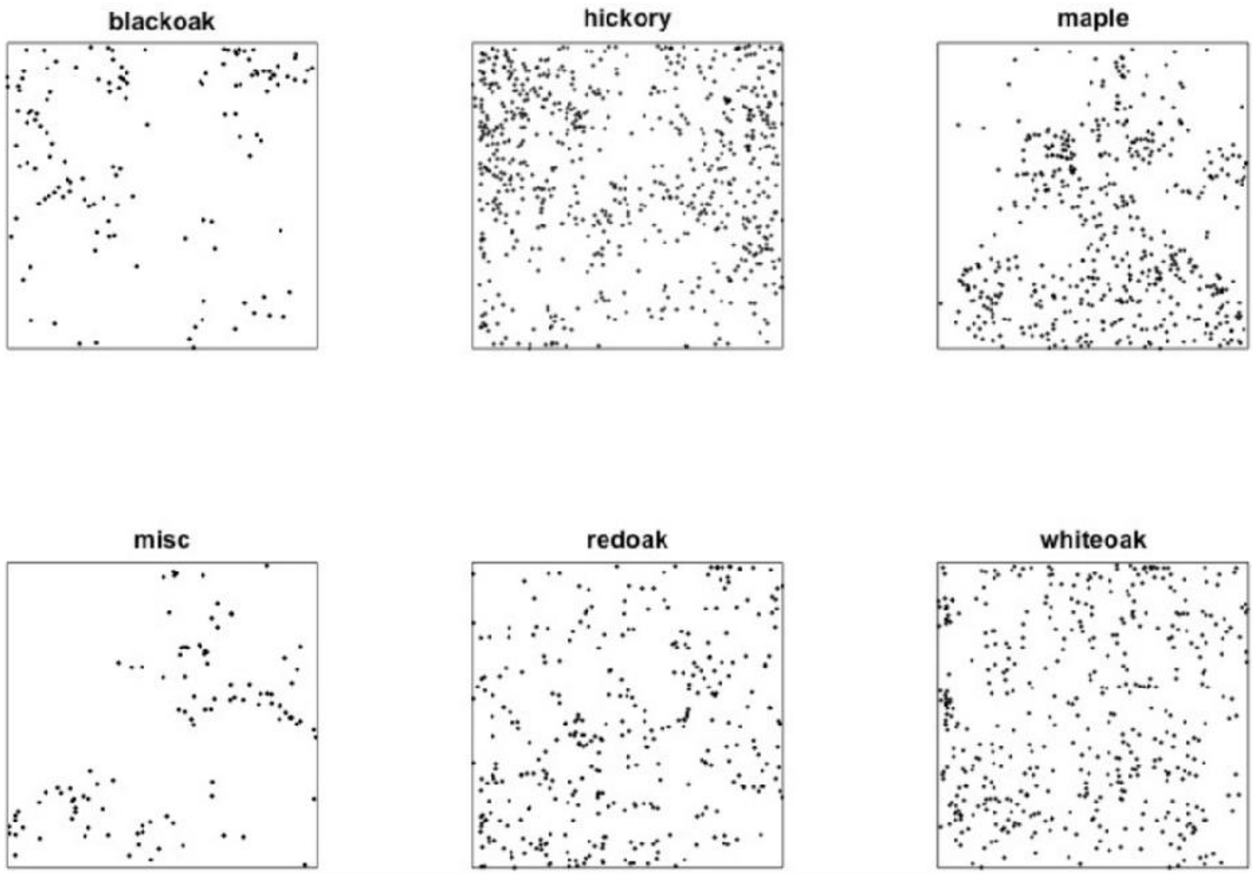
Some requirements

- Data mapped on a 2D isotropic plane
- Study area boundary determined objectively (MAUP)
- Population rather than sample data – one to one correspondence between points and objects/events in the real world
- Data can be just the locations, or locations with some kind of attribute value ⇢ marked pattern

Point patterns: marked vs unmarked

categorical mark





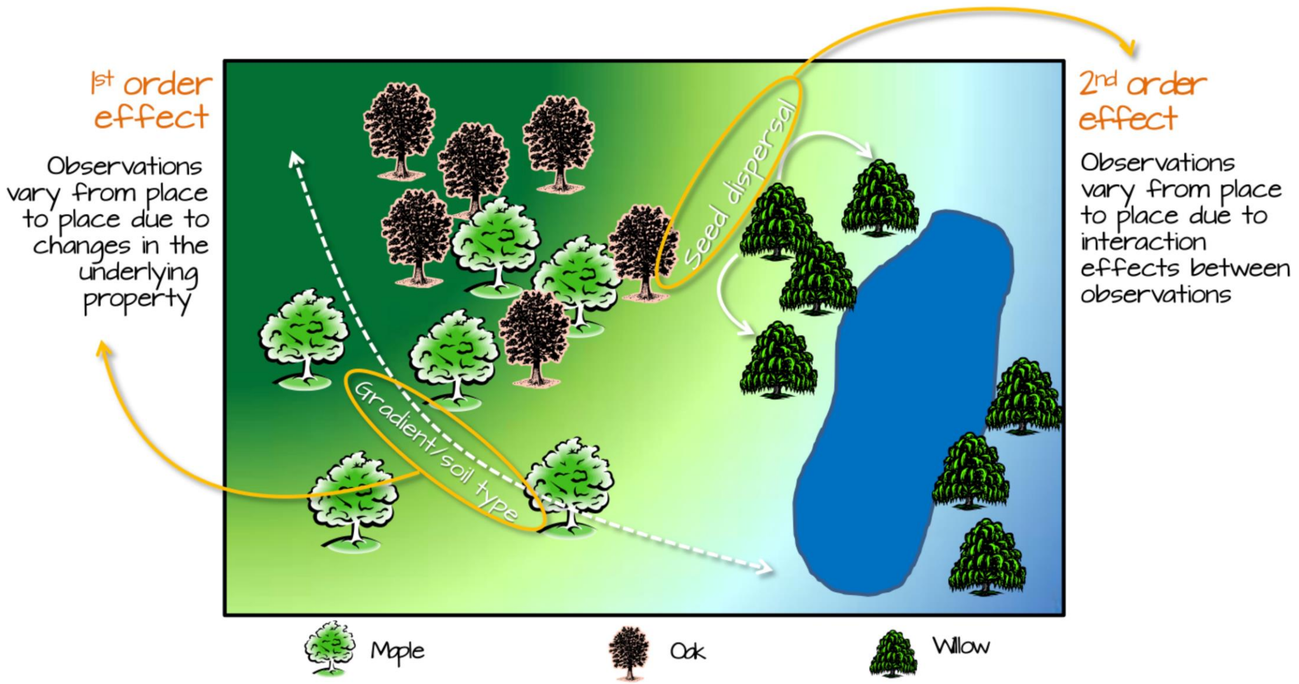
Analysing point patterns: First- or second-order?

First order moment

- Intensity of a process across space
- Probability that an event occurs within a region
- **Absolute** location is important

Second order moment

- Interaction between locations, depending on distance
- Probability that two events occur within a certain distance of each other
- **Relative** location is important



Analysing point patterns: First- or second-order?

Examples

- How many Airbnbs are there per SqKM in each of the Copenhagen neighborhoods?
- What is the chance of getting into an accident when cycling in a given city?
- How many reported crimes are there for each municipality?

Analysing point patterns: First- or second-order?

Examples

- Are new Covid-19 incidents more likely to be found in the vicinity of existing ones?
- Can we identify clusters of cycling accidents?
- Is there a minimum distance between any two human settlements (dispersion)?

Point pattern analysis methods

Method	Description	
Points	Exploratory Data Analysis Measuring geographic distributions • Mean Center; Central/Median Center • Standard Distance; • Standard Deviation/Standard Deviational Ellipse	**Summary statistics**
Kernel Density Estimate	Exploratory Data Analysis Is an example of "exploratory spatial data analysis" (ESDA) that is used to "visually enhance" a point pattern by showing where features are concentrated.	
Quadrat analysis	Exploratory Data Analysis Measuring intensity based on density (or mean number of events) in a specified area. Calculate variance/mean ratio	**Density-based approaches**
Nearest neighbor analysis	Distance-based Analysis Measures spatial dependence based on the distances of points from one another. Calculates a nearest neighbor index based on the average distance from each feature to its nearest neighboring feature.	
Ripley's K	Distance-based Analysis Measures spatial dependence based on distances of points from one another. K (d) is the average density of points at each distance (d), divided by the average density of points in the entire area.	**Distance-based approaches**

🔗 *https://www.e-education.psu.edu/geog586/node/669*

Applications of point pattern analysis

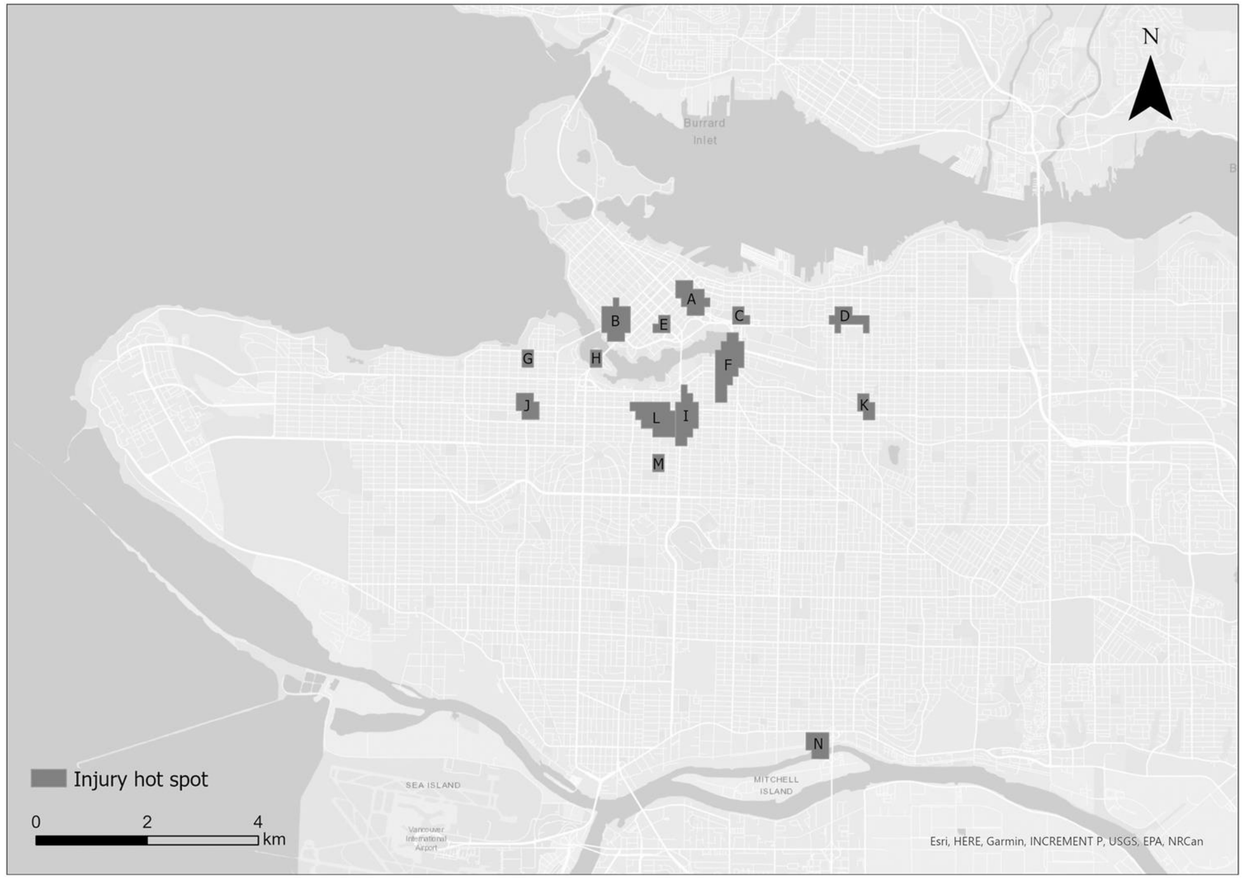
Cholera outbreak, London – 1854

Dr. John Snow mapped the cholera cases and was able to track down a water pump as the source. The pump was shut and the outbreak stopped.

Wikipedia: 1854 Broad Street cholera outbreak

Applications of point pattern analysis

Bicycling incidents hot spots – Vancouver



The authors converted bicycling incidents to hot spots and grouped the narrative data by hot spot, type of incident, and type of object involved in the incident.

- Kernel Density Estimation (KDE)

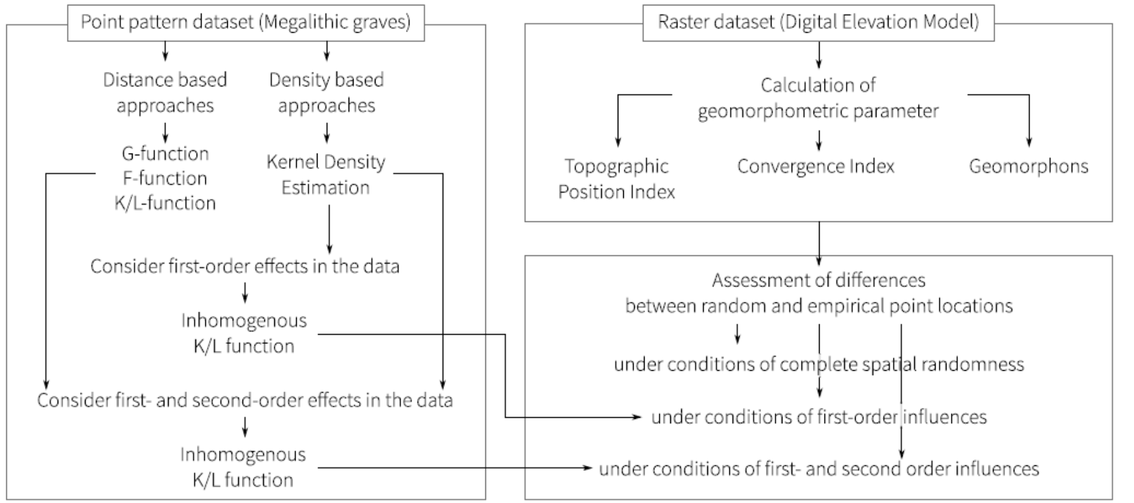
J. Fischer et al., "Spatial variation in bicycling risk based on crowdsourced safety data," Can. Geogr., vol. 66, no. 3, pp. 556–568, 2022, doi: 10.1111/cag.12756.

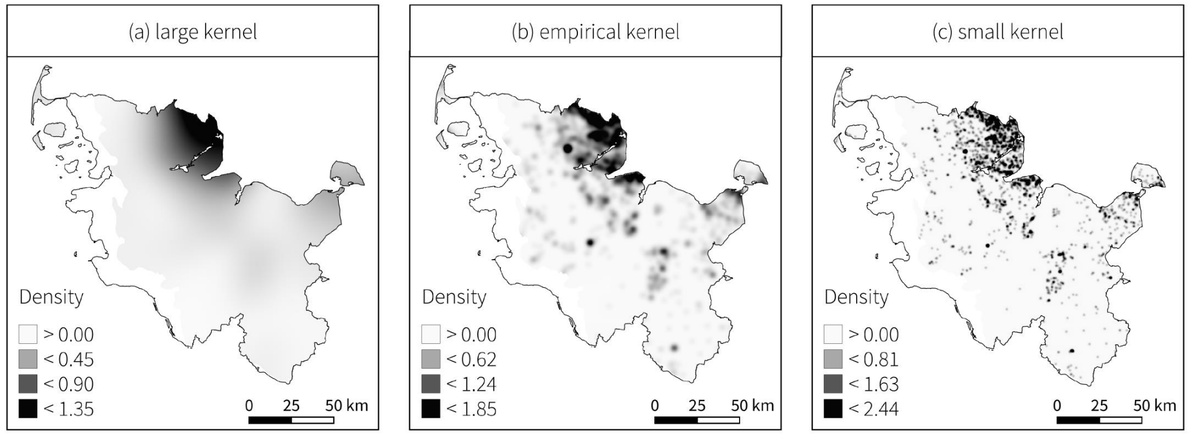
Applications of point pattern analysis

PPA in Digital Geoarcheology

The authors here had the locations of megalithic graves from the bronze age and they applied methods of ppa to reconstruct the spatial processes that created the sample

- Density based measures → the influence of first order effects on the dataset
- Distance related approaches →investigate the interaction between the points



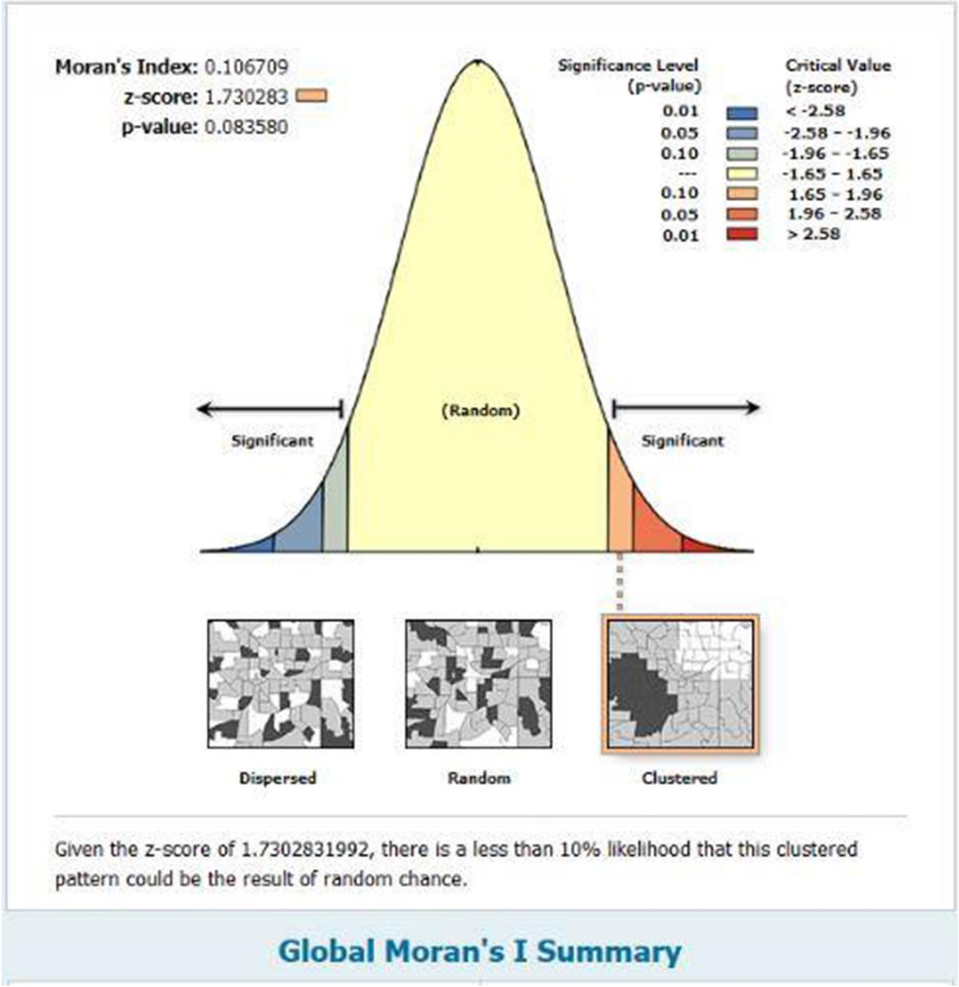


D. Knitter and O. Nakoinz, "Point Pattern Analysis as Tool for Digital Geoarchaeology: A Case Study of Megalithic Graves in Schleswig-Holstein, Germany," pp. 45–64, 2018, doi: 10.1007/978-3-319-25316-9_4.

Applications of point pattern analysis

PPA for Covid-19 outbreak

The authors examined the spread of Covid-19 cases in Qom Province, Iran, in one day. The calculated Moran's Index and the Z-score, indicating high clustering in some areas

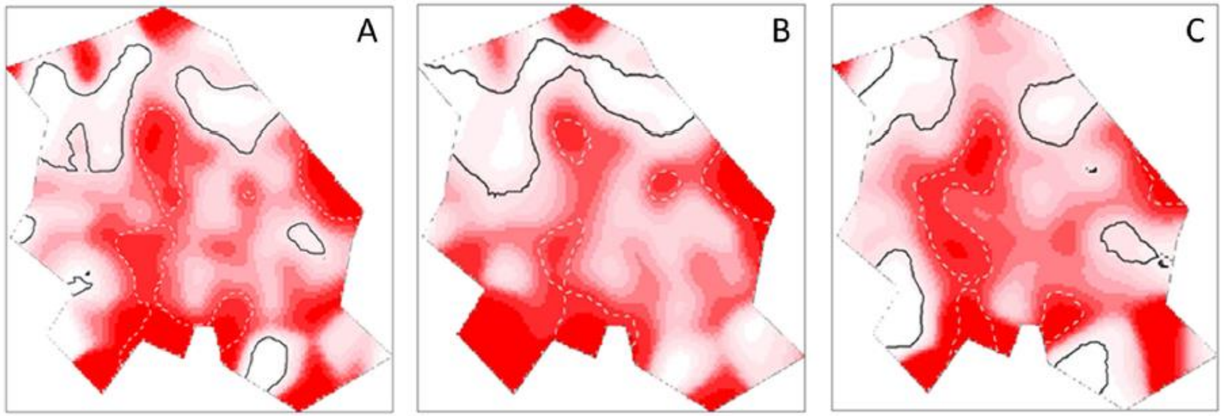


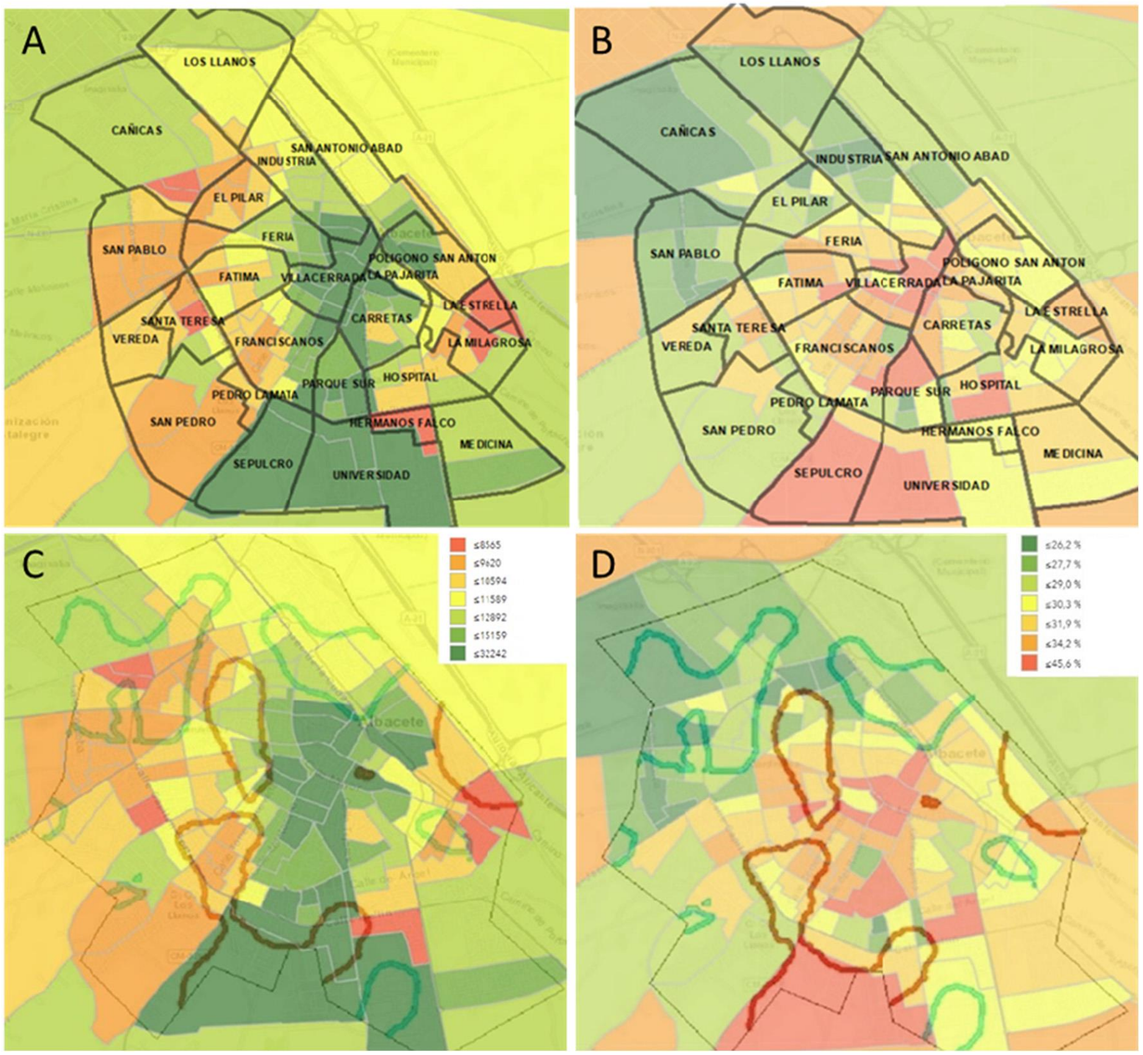
N. Jesri et al., "Mapping and Spatial Pattern Analysis of COVID-19 in Central Iran Using the Local Indicators of Spatial Association (LISA)," BMC Public Health, vol. 21, no. 1, pp. 1–10, 2021, doi: 10.1186/s12889-021-12267-6.

Applications of point pattern analysis

PPA for Covid-19 outbreak (2)

The authors proposed using spatial pp to delimit spatial units of analysis based on the highest local incidence of hospitalizations instead of administrative limits.





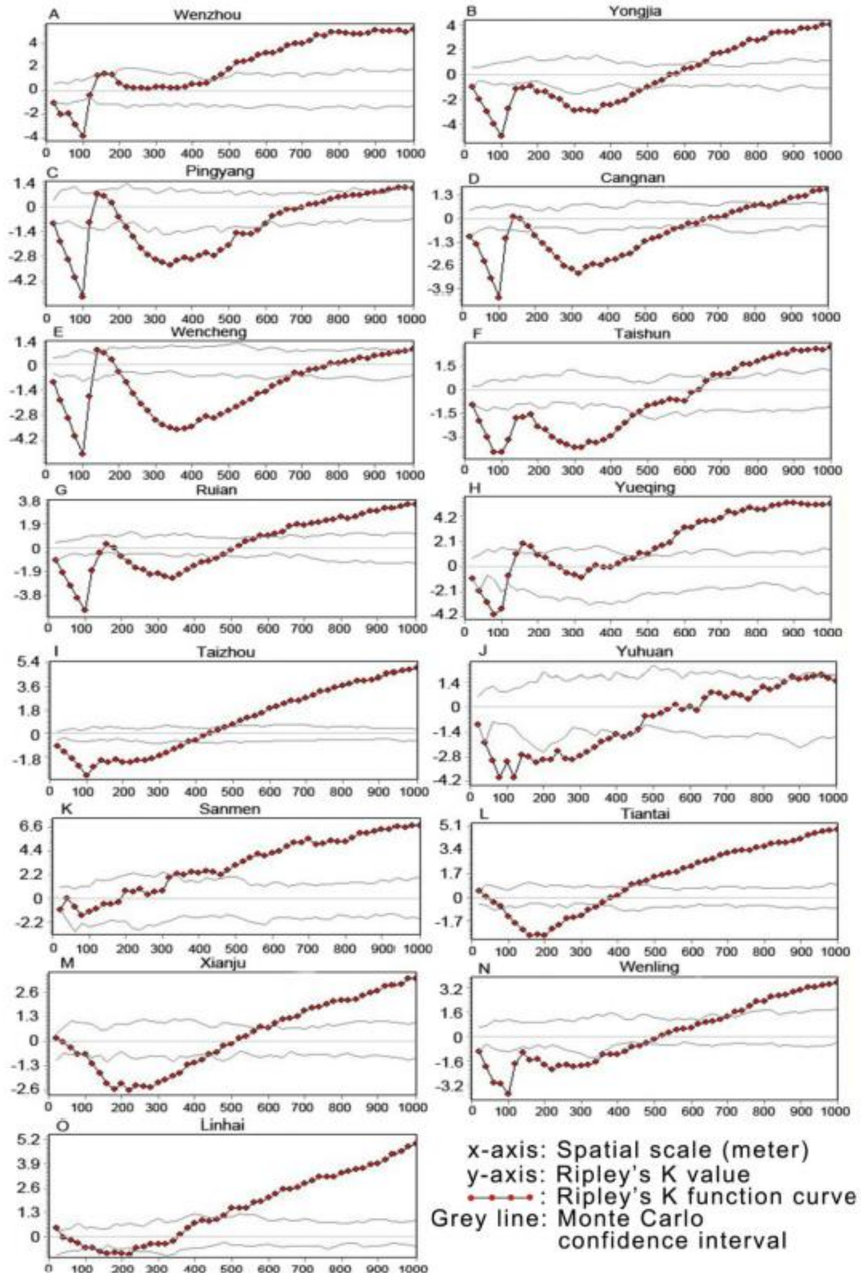
M. Garcia-Morata, J. Gonzalez-Rubio, T. Segura, and A. Najera, "Spatial analysis of COVID-19 hospitalised cases in an entire city: The risk of studying only lattice data," Sci. Total Environ., vol. 806, p. 150521, 2022, doi: 10.1016/j.scitotenv.2021.150521.

Applications of point pattern analysis

PPA of human settlements

The authors used Ripley's K function and Monte Carlo simulation to investigate human settlement point patterns. Results indicated that human settlements displayed regular-random-cluster patterns from small to big scale

Figure 3. Ripley's K point pattern of human settlements in Wen-Tai region, China. Figures show the Ripley's K function curve (grey line with red dots) between the two grey lines as the Monte Carlo confidence interval in different counties.

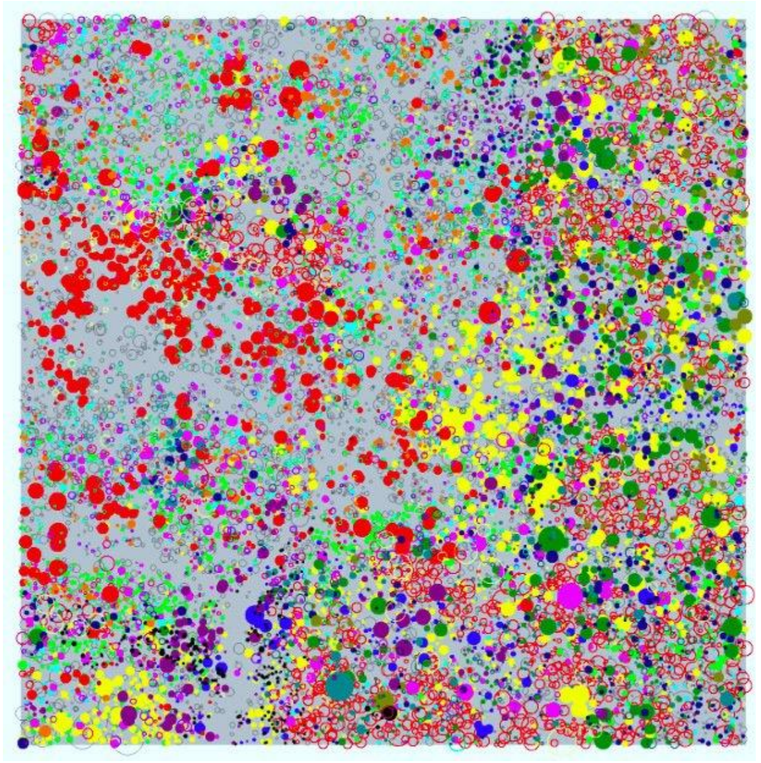


Z. Zhang, R. Xiao, A. Shortridge, and J. Wu, "Spatial point pattern analysis of human settlements and geographical associations in eastern coastal China - A case study," Int. J. Environ. Res. Public Health, vol. 11, no. 3, pp. 2818–2833, 2014, doi: 10.3390/ijerph110302818.

Applications of point pattern analysis

Programita

The software Programita was developed for conducting point-pattern analysis with several summary functions. Programita contains many standard and non-standard null models that cover most practical applications of point pattern analysis in ecology.



Spatial Point Pattern Analysis

Point pattern analysis methods

Method	Description
Points	Exploratory Data Analysis Measuring geographic distributions • **Mean Center**; Central/Median Center • **Standard Distance**; • Standard Deviation/Standard Deviational Ellipse
Kernel Density Estimate	Exploratory Data Analysis Is an example of "exploratory spatial data analysis" (ESDA) that is used to "visually enhance" a point pattern by showing where features are concentrated.
Quadrat analysis	Exploratory Data Analysis Measuring intensity based on density (or mean number of events) in a specified area. Calculate variance/mean ratio
Nearest neighbor analysis	Distance-based Analysis Measures spatial dependence based on the distances of points from one another. Calculates a nearest neighbor index based on the average distance from each feature to its nearest neighboring feature.
Ripley's K	Distance-based Analysis Measures spatial dependence based on distances of points from one another. K (d) is the average density of points at each distance (d), divided by the average density of points in the entire area.

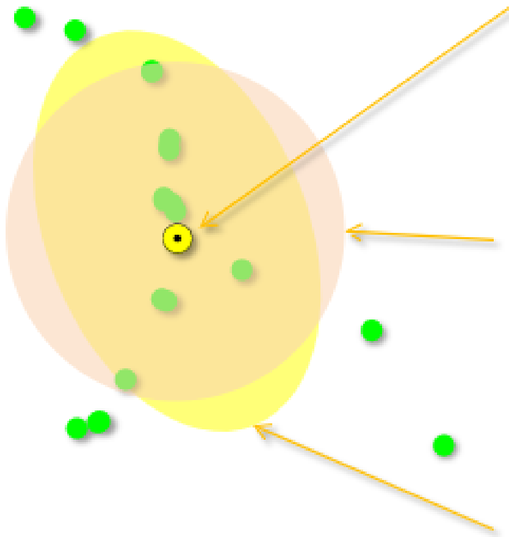
Summary statistics

Density-based approaches

Distance-based approaches

🔗 *https://www.e-education.psu.edu/geog586/node/669*

Summary statistics and centrography



Mean center
computed average X and Y coordinate values.

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

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

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

Standard deviational ellipse
separate standard distances for each axis.

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

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

---

https://mgimond.github.io/Spatial/chp11_0.html

Mean center

- Mean center averages the X and Y coordinates of all points in the study area
  (⇢ first order moment).

$$\overline{s} = (\mu_x, \mu_y,) = (\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n})$$

or

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

Mean center
computed average X
and Y coordinate
values.

Where (μ_x, μ_y) are the coordinates of the mean center, (x_i, y_i) represent the coordinates of a given point i, and n is the total number of points.

Standard distance

- Similar to standard deviation
- Measures how far each event is away from the mean center, on average:

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

Where (μ_x, μ_y) are the coordinates of the mean center, (x_i, y_i) represent the coordinates of a given point ***i***, and ***n*** is the total number of points.

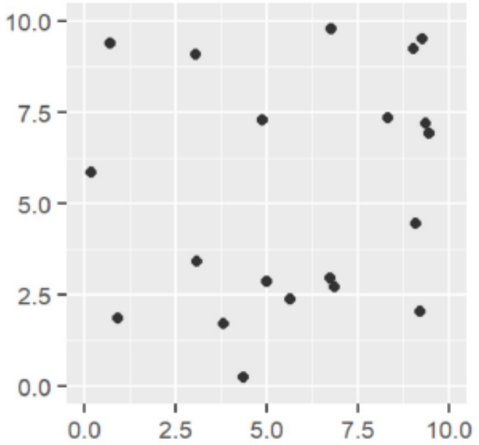
In combination, the mean centre and the standard distance can be used to plot a summary circle.

Density

- Overall **intensity** of a point pattern
- **Global** measures are only suitable when study areas are small

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

A basic measure of a pattern's density **λ** is its overall, or **global**, density. This is simply the ratio of the observed number of points, **n**, to the study region's surface area, **a**.



An example of a point pattern where n = 20 and the study area (defined by a square boundary) is 10 units squared. The point density is thus 20/100 = 0.2 points per unit area.

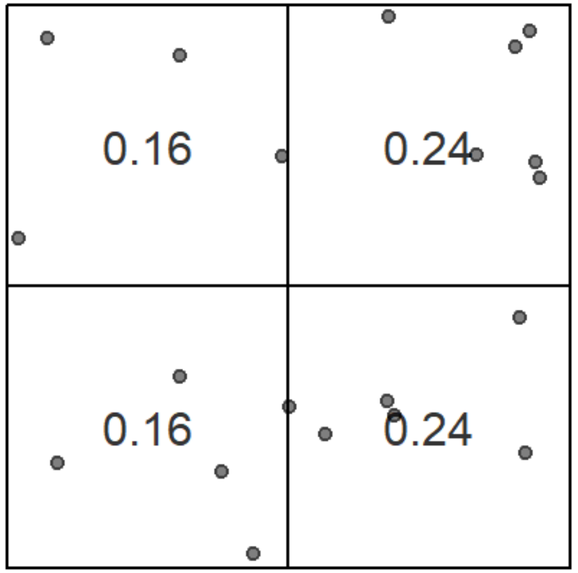
Density-based approaches

- Overall **intensity** of a point pattern
- Global measures are only suitable when study areas are small
- ***Local*** measures quantify variability throughout a study area

    - *Quadrat density*
    - *Kernel density*

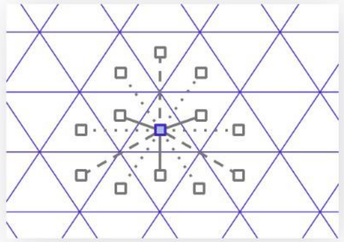
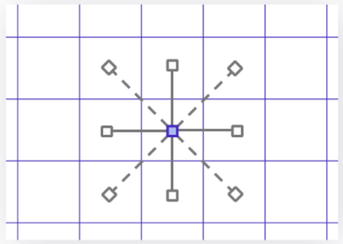
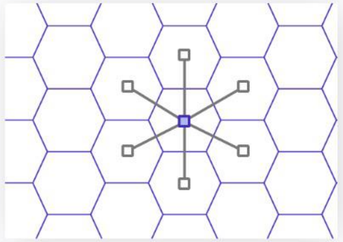
Local density – Quadrat density

- Place quadrats (cells of a fixed size and shape) over a study area
- Quadrats can be either exhaustive or random
- Count events in each quadrat
- The result is a frequency distribution of counts that can be compared to an expected distribution, i.e. analyse differences between observed and expected counts



Local density – Quadrat density

- The cell shape of the grid system is an important consideration.
- For simplicity, it should be a polygon that tiles regularly: the triangle, the square, or the hexagon.
- Hexagons, triangles and squares have neighbors with different distances.
- We often use a regular shaped quadrats and expect the same number of events (points) in each quadrat.

Triangle	Square	Hexagon
		
Triangles have 12 neighbors	Squares have 8 neighbors	Hexagons have 6 neighbors

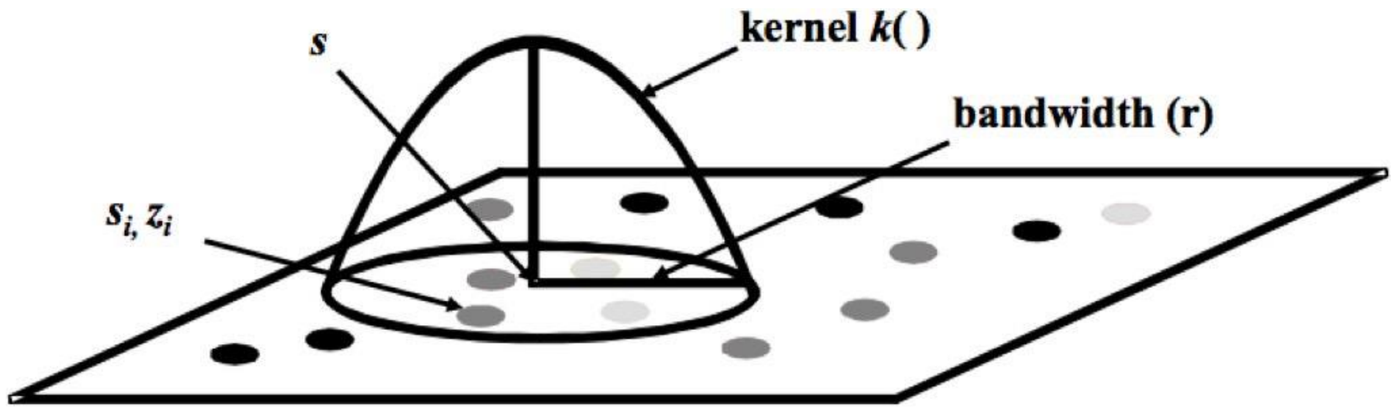
🔗 *https://www.uber.com/en-SE/blog/h3/*

Local density – Kernel density

- Pattern has a density at any location, not just at locations where there is an event.
- Density is continuous

The principle

- Volume under is equal to the value measured at location
- Distributes this value over a circular area with radius using kernel function

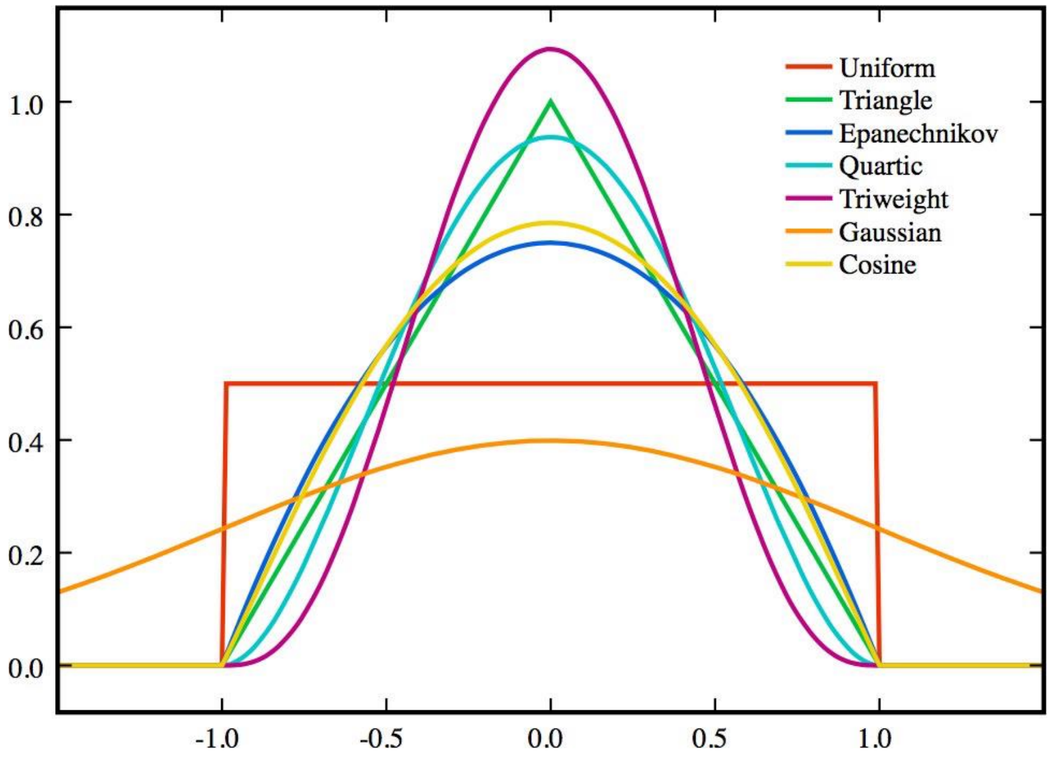


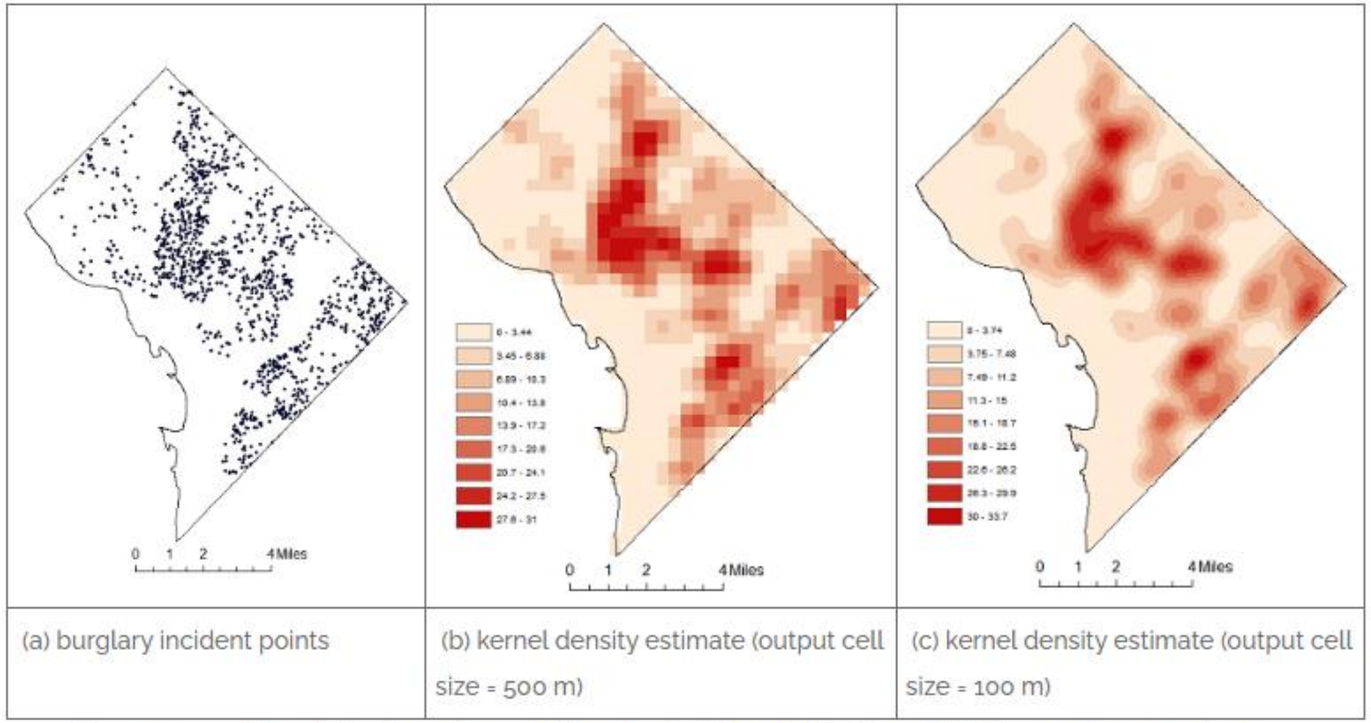
---

https://www.uber.com/en-SE/blog/h3/

Kernel density: functions and bandwidth

- Selecting the suitable smoothness for the case





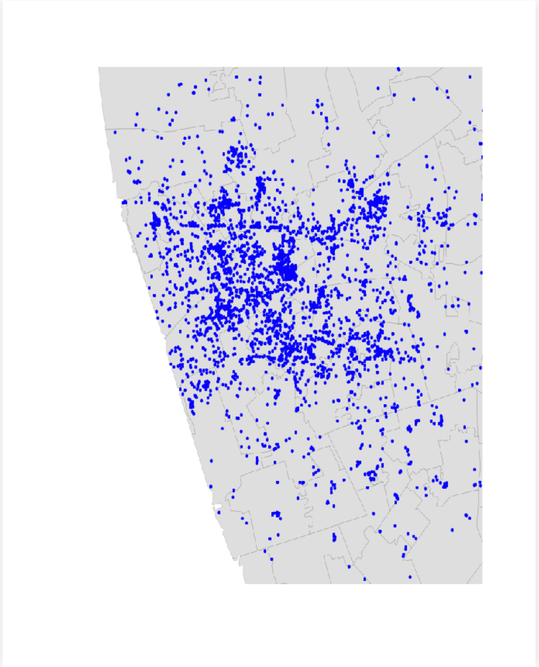
(a) burglary incident points | (b) kernel density estimate (output cell size = 500 m) | (c) kernel density estimate (output cell size = 100 m)

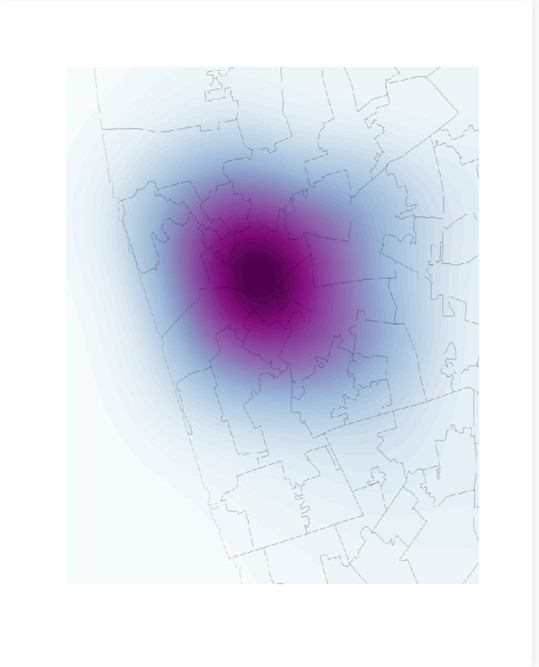
Figure 2a - 2c. Kernel density estimates of burglary incidents in Washington, D.C., 2018. Source: author.

- https://en.wikipedia.org/wiki/Kernel_(statistics)
- *https://gistbok.ucgis.org/bok-topics/kernels-and-density-estimation*

Local density – Kernel density benefits

- Improved visualization
- Map of estimates of locations (e.g. showing estimated risk for bike accidents per cell)
- A first-order stationary process should only show location variations from the average intensity, rather than marked trends across the study region
- Solves MAUP
- Allows point objects to be linked with other geographic (spatially continuous) data





Distance-based approaches

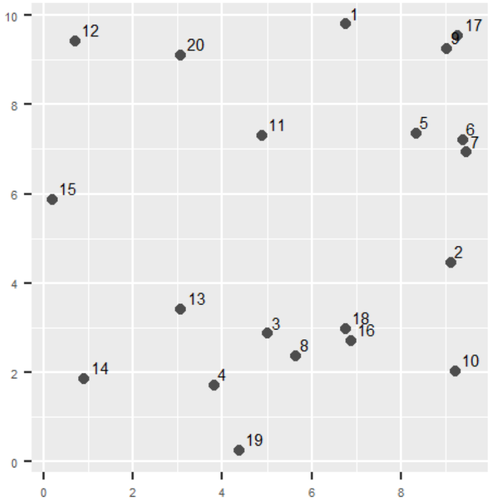
- How the points are distributed relative to one another (⇢ second order moment)

- And not how the points are distributed relative to the study extent (density-based approaches).

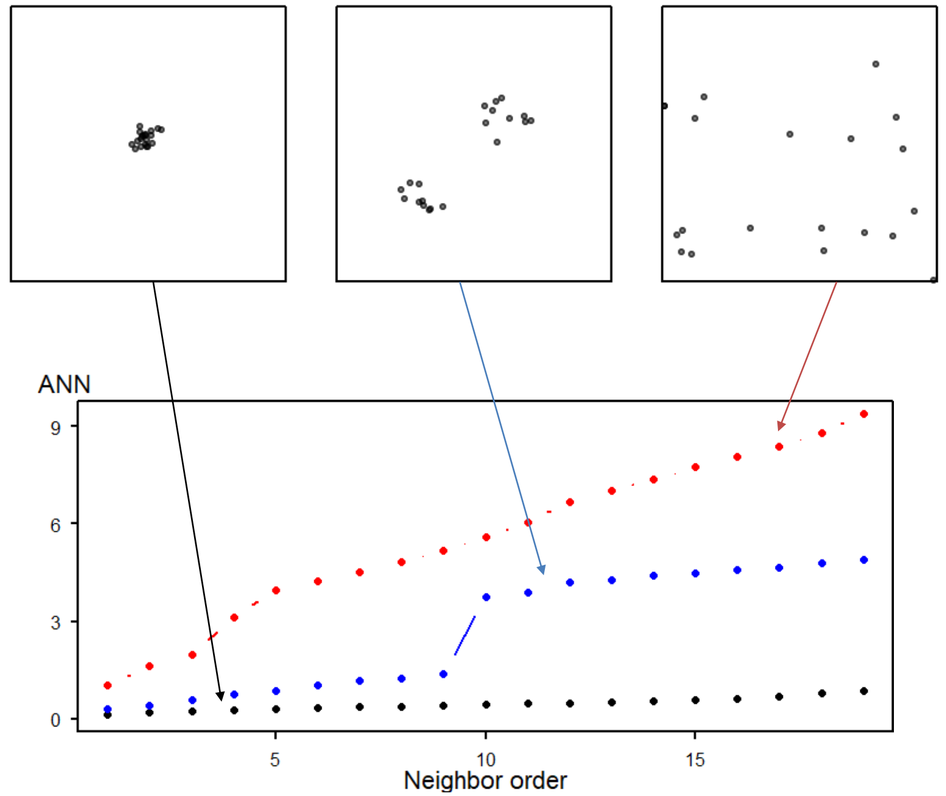
    - Average nearest neighbor (ANN)
    - Distance functions: G, F, K functions

Distance-based approaches – ANN

- Calculate Euclidian distance between each point and its nearest neighbor
- Interpretation:
  NND < 1: clustering
  NND > 1: dispersion
- Quadrat test depends on the size of quadrats, NND solves this problem

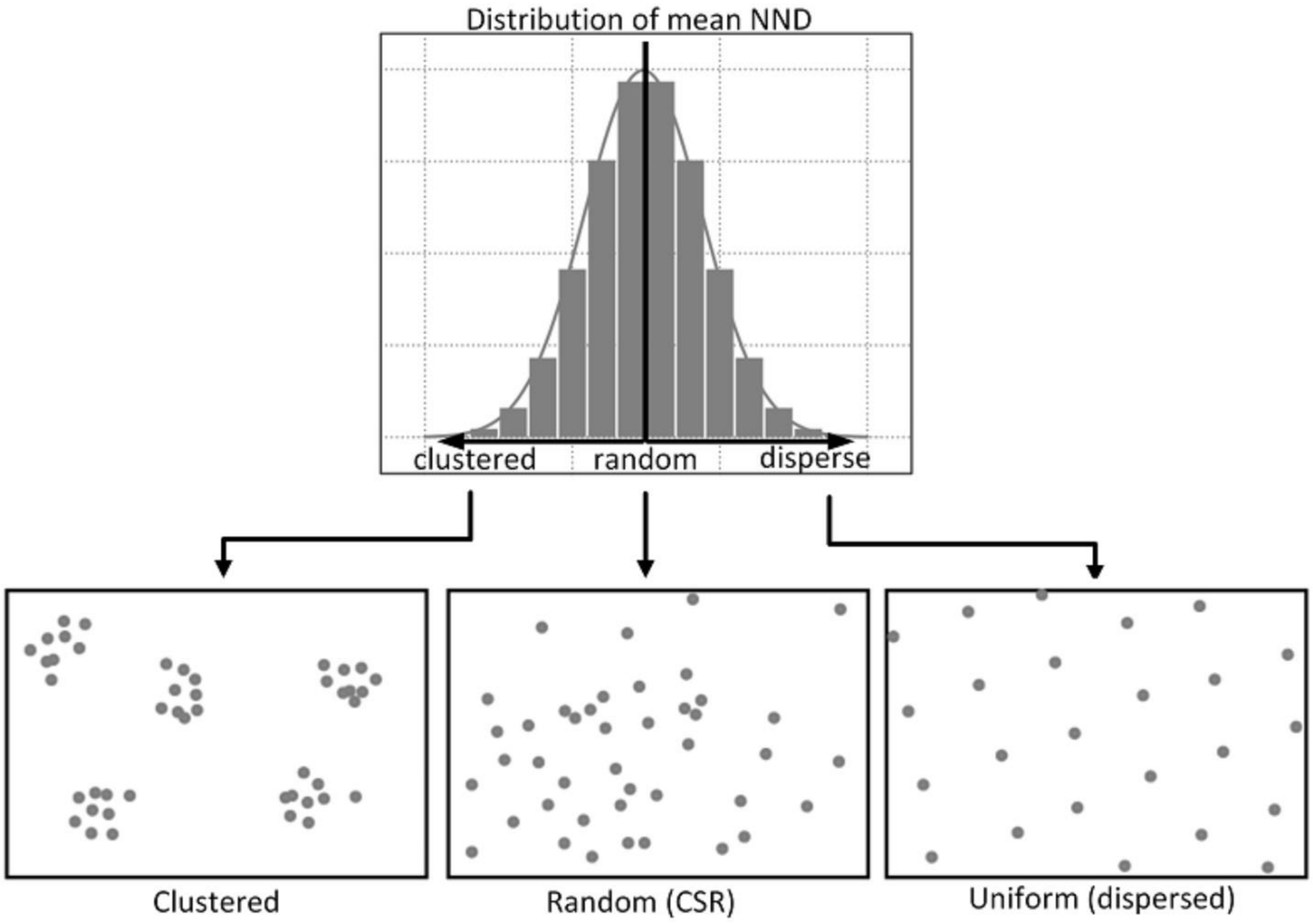


From	To	Distance	From	To	Distance
1	9	2.32	11	20	2.55
2	10	2.43	12	20	2.39
3	8	0.81	13	4	1.85
4	19	1.56	14	13	2.67
5	6	1.05	15	12	3.58
6	7	0.3	16	18	0.29
7	6	0.3	17	9	0.37
8	3	0.81	18	16	0.29
9	17	0.37	19	4	1.56
10	2	2.43	20	12	2.39



https://mgimond.github.io/Spatial/chp11_0.html#distance-based-analysis

Nearest Neighbor Distance (NND)



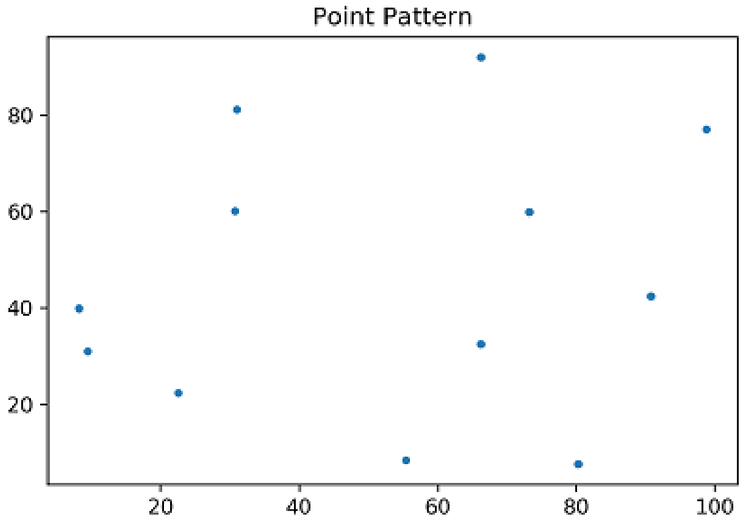
https://gistbok.ucgis.org/bok-topics/point-pattern-analysis

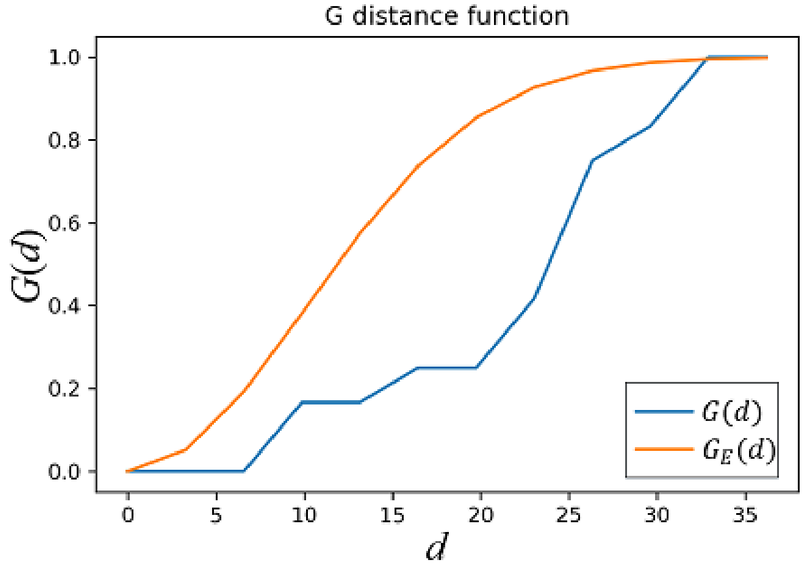
Distance-based approaches – G function

- Describes variations of a pp
- Is the fraction of points which have a neighbor within distance ***d***
- Measures clustering for a distance ***d***
- Calculates the cumulative frequency distribution of the NND of a pp

$$G(d) = \frac{sum(D_{ij} < d)}{n} \quad (12)$$

where $sum(D_{ij} < d)$ stands for the number of point pairs i and j with a distance smaller than d, and n represents the total number of points (Figure 4).

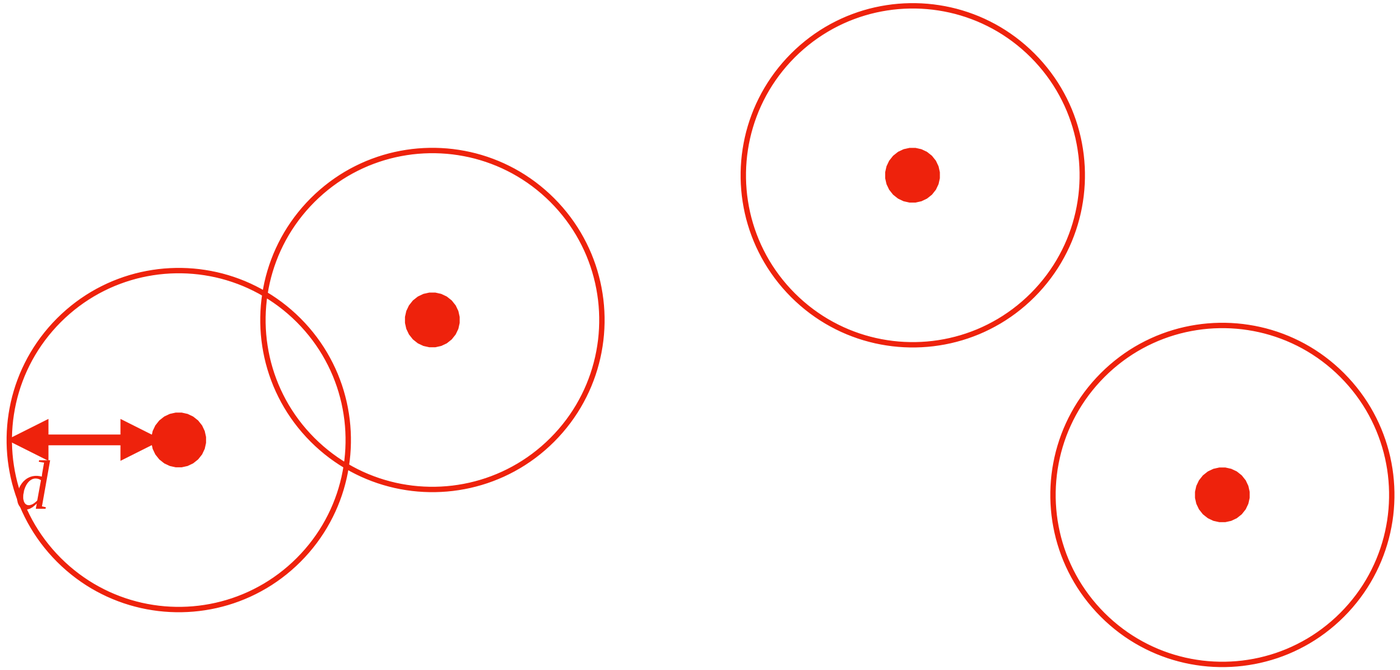




G function

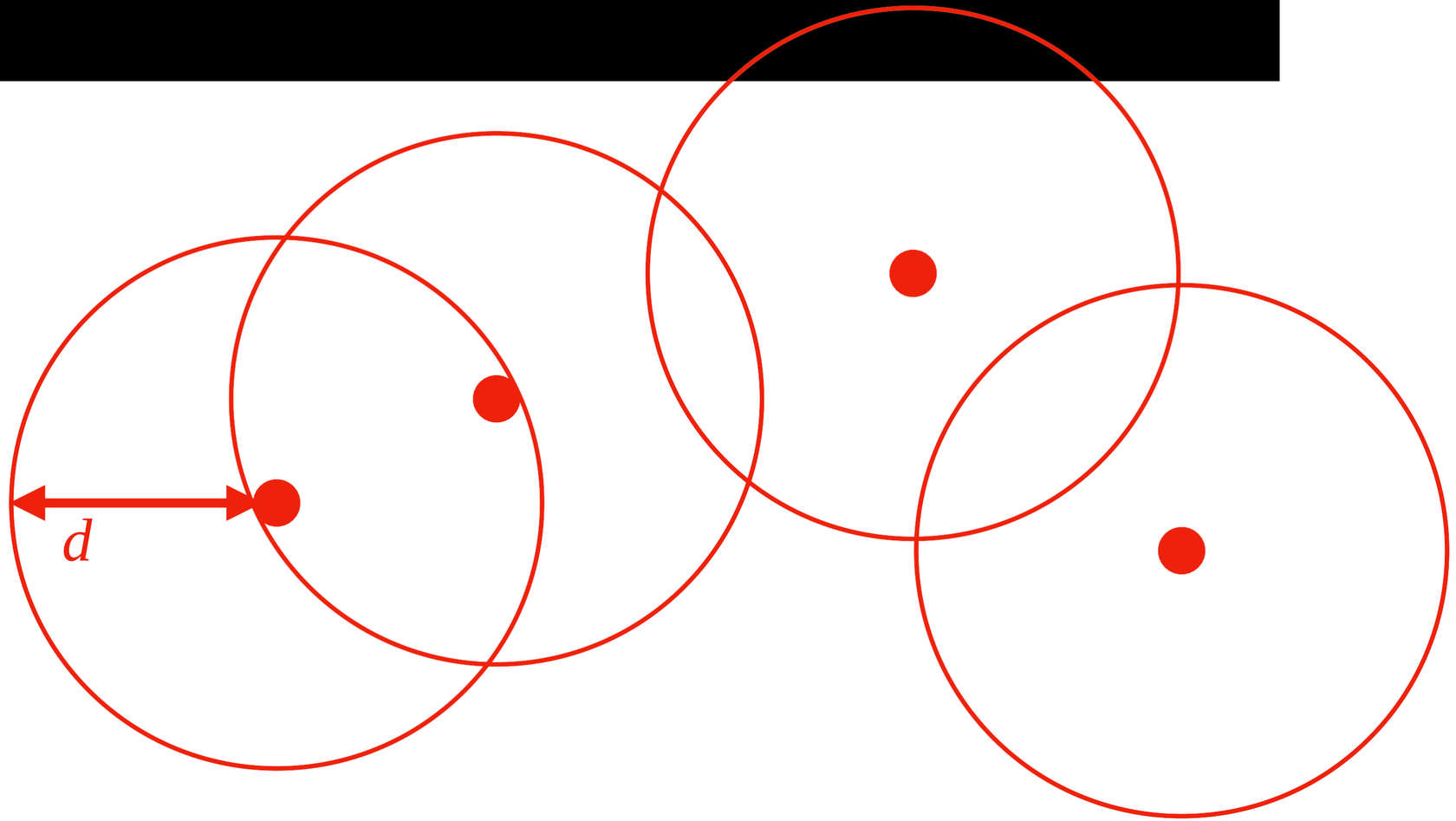
Data points

G function



$$G(1) = 0/4 = 0$$

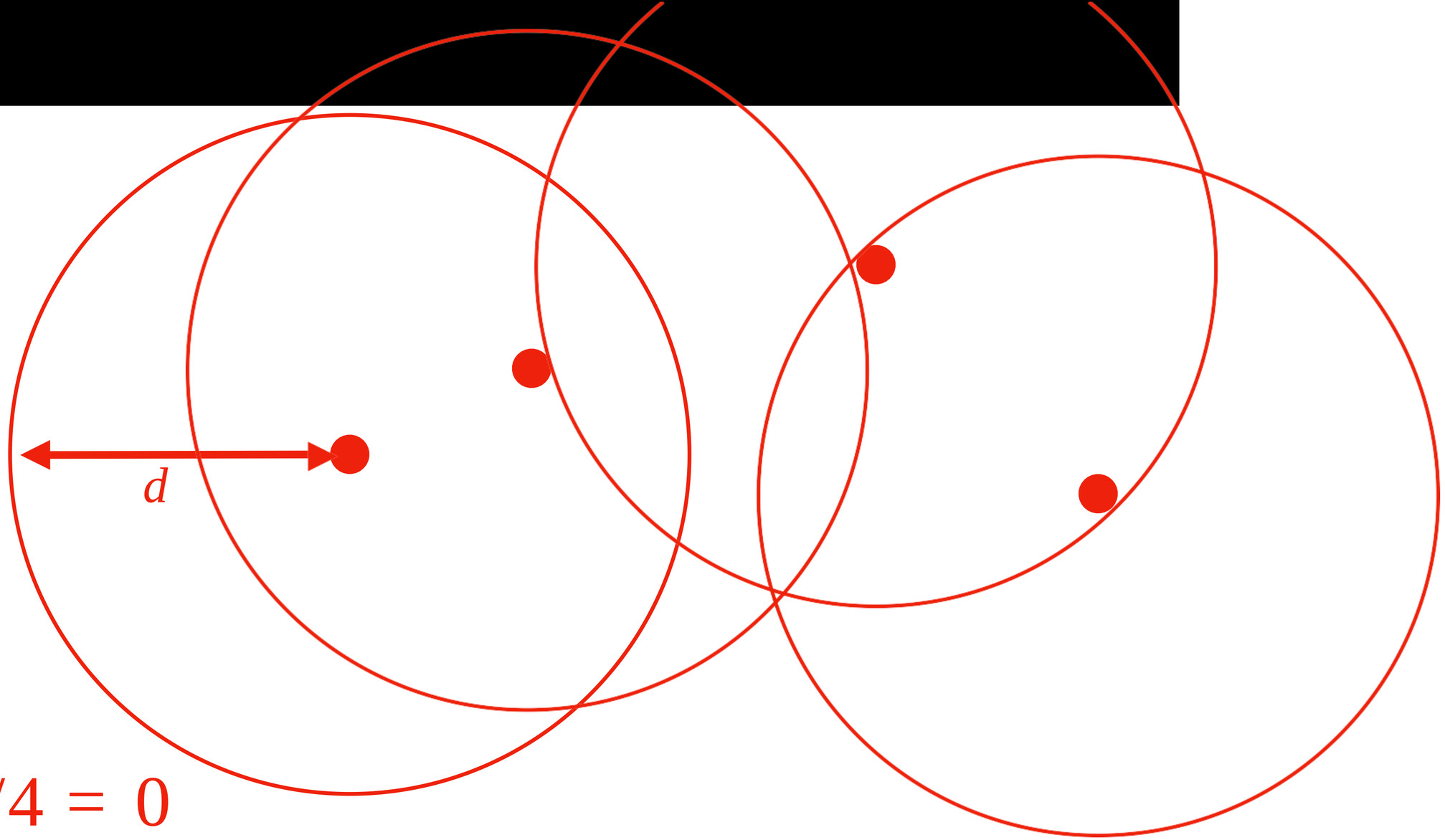
G function



$G(1) = 0/4 = 0$

$G(2) = 2/4 = 0.5$

G function



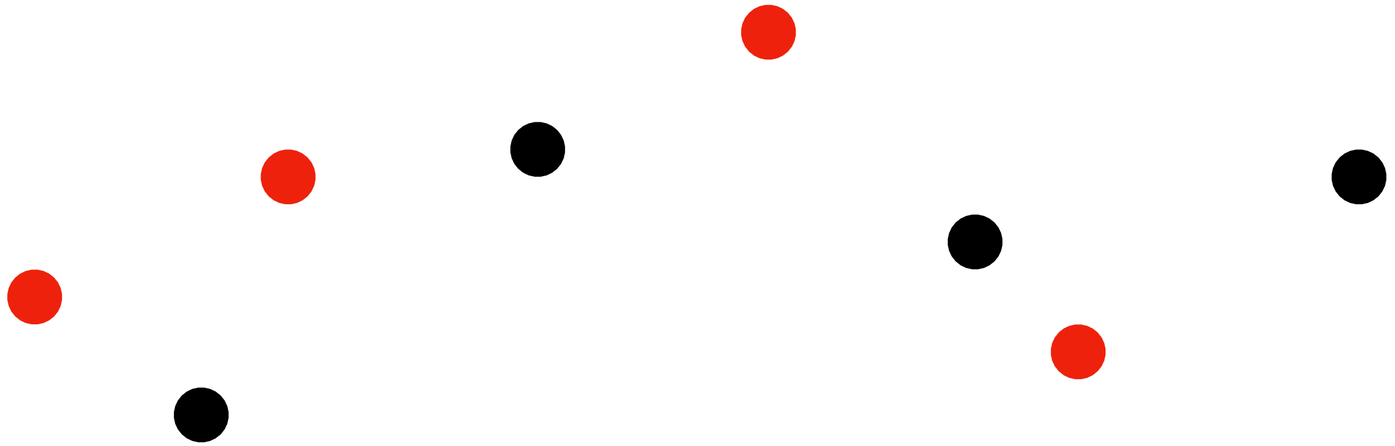
$G(1) = 0/4 = 0$

$G(2) = 2/4 = 0.5$

$G(3) = 4/4 = 1$

G function

Now compare it to random points

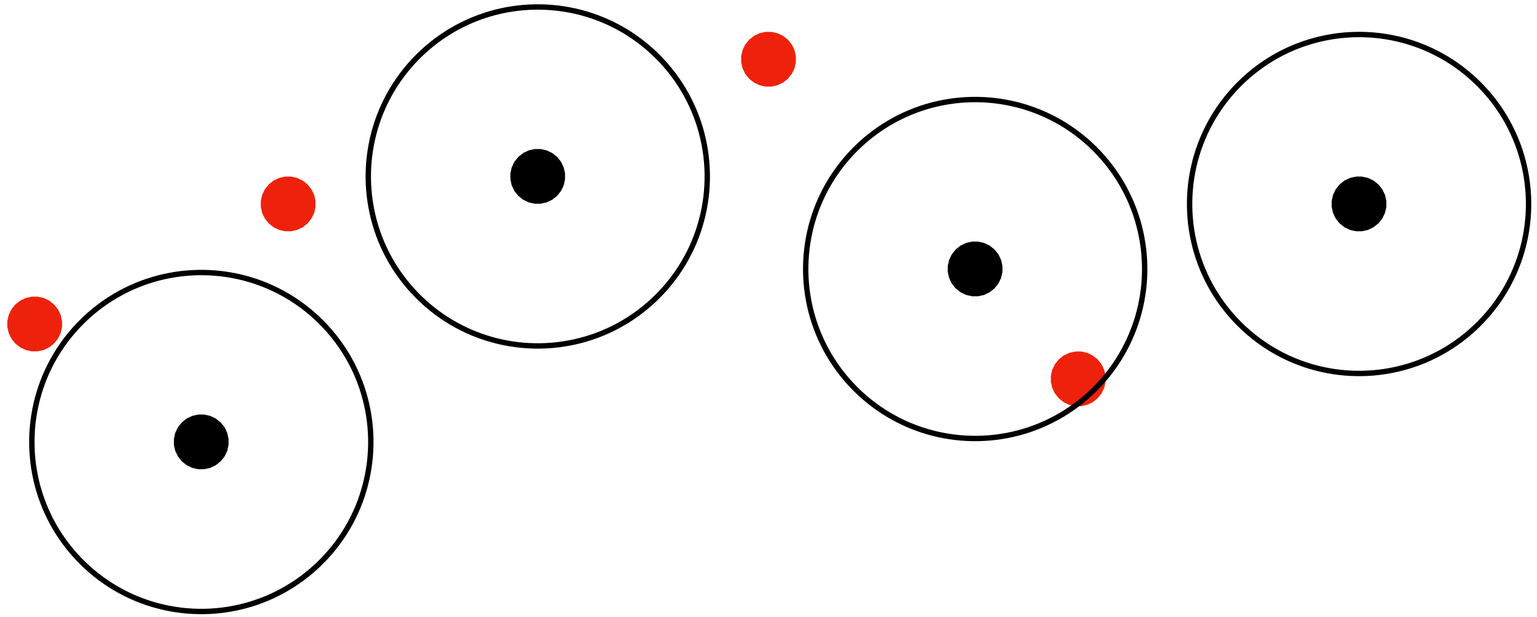


$G(1) = 0/4 = 0$

$G(2) = 2/4 = 0.5$

$G(3) = 4/4 =$

G function measures clustering for a distance *d*



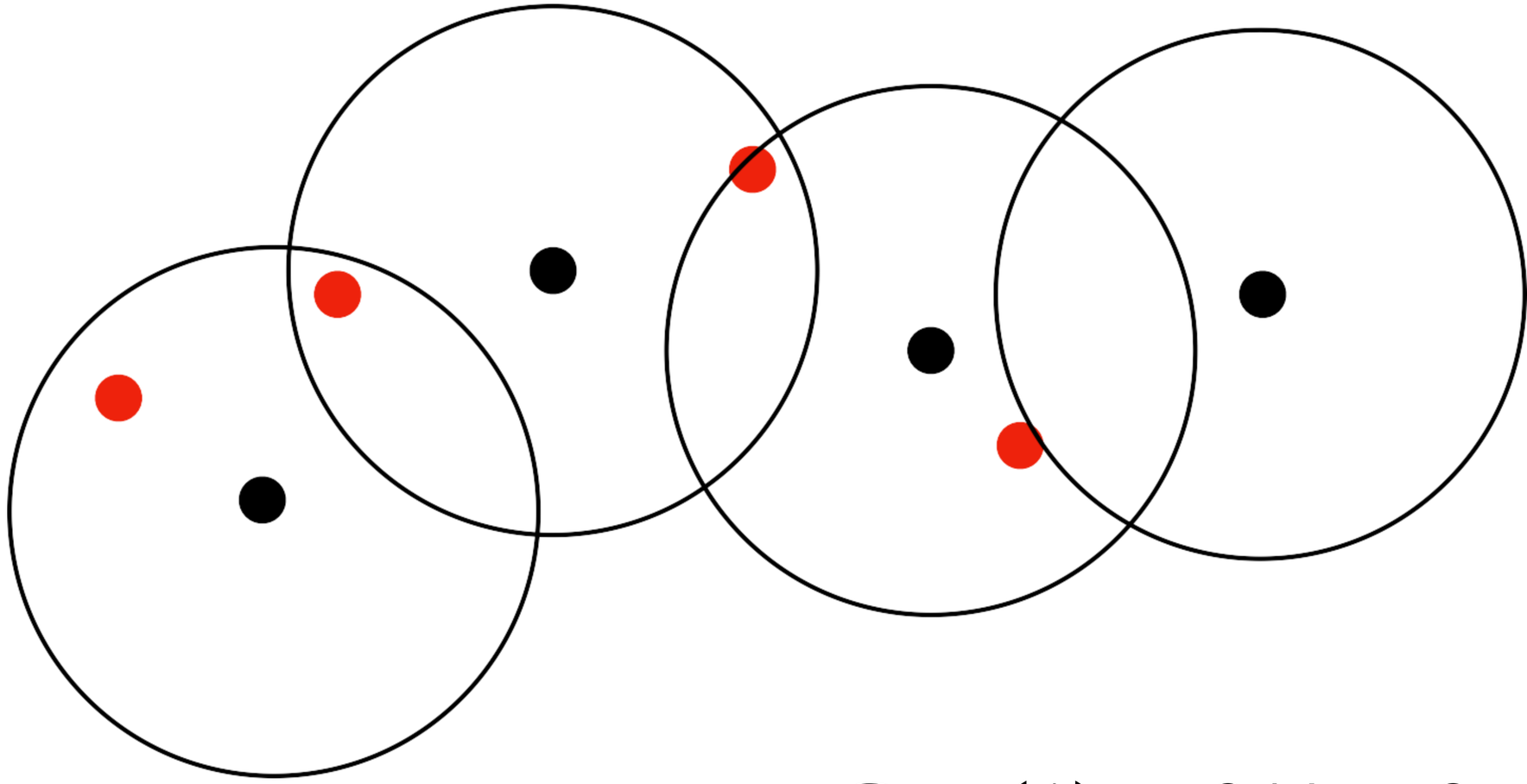
$G(1) = 0/4 = 0$

$G(2) = 2/4 = 0.5$

$G(3) = 4/4 =$

$G_{rand}(1) = 0/4 = 0$

G function measures clustering for a distance *d*



$G(1) = 0/4 = 0$

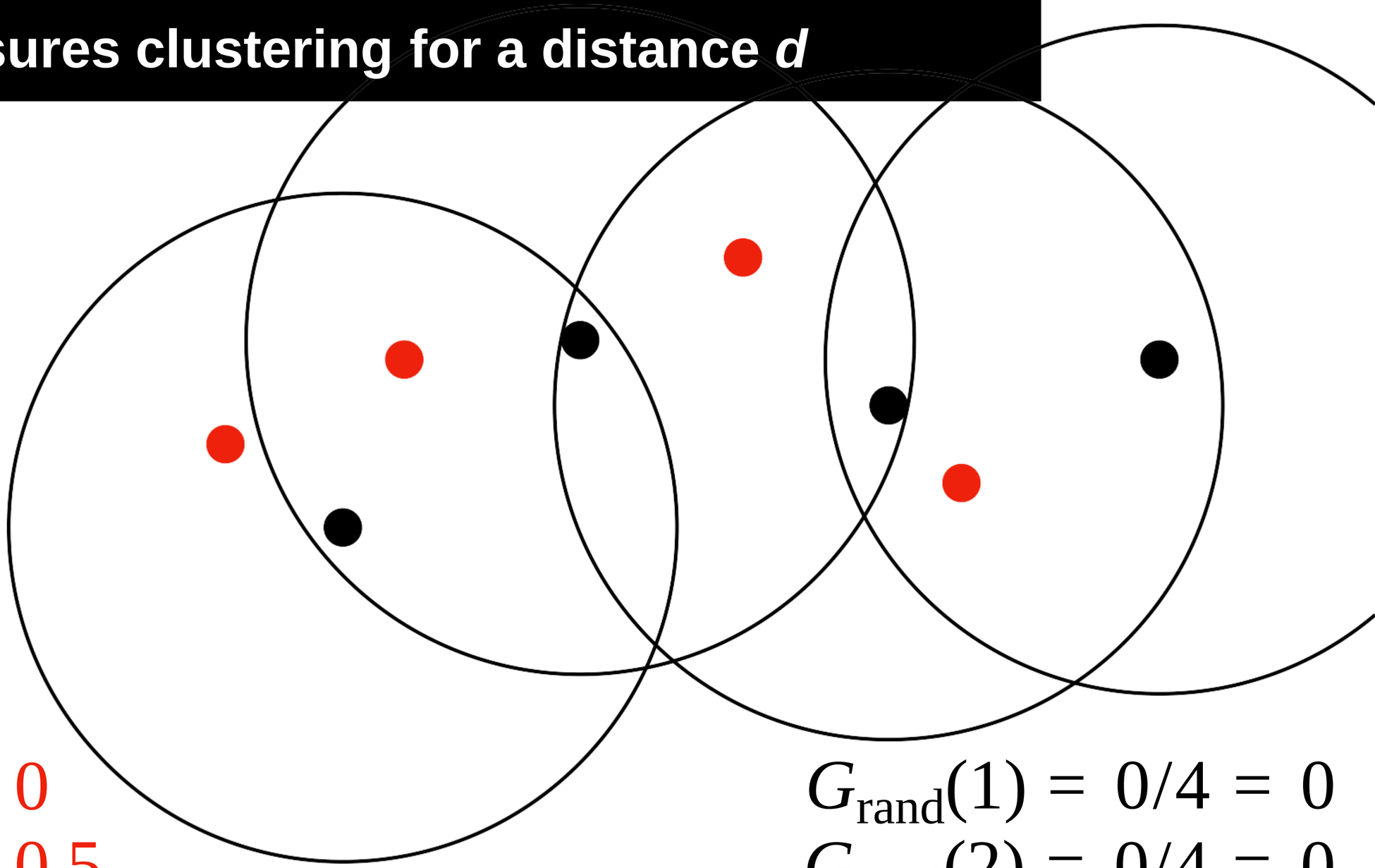
$G(2) = 2/4 = 0.5$

$G(3) = 4/4 =$

$G_{rand}(1) = 0/4 = 0$

$G_{rand}(2) = 0/4 = 0$

G function measures clustering for a distance *d*



$G(1) = 0/4 = 0$

$G(2) = 2/4 = 0.5$

$G(3) = 4/4 =$

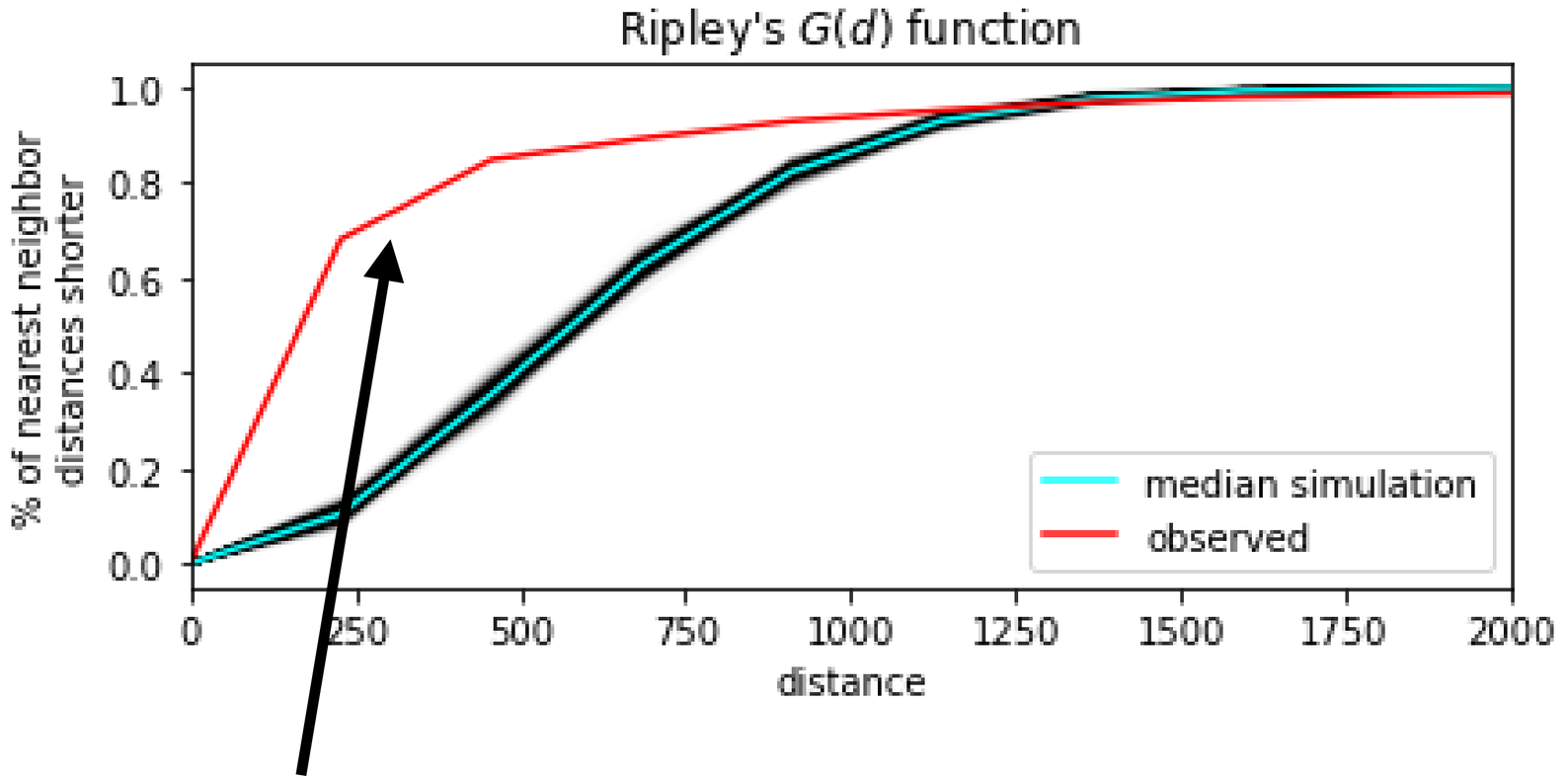
$G_{rand}(1) = 0/4 = 0$

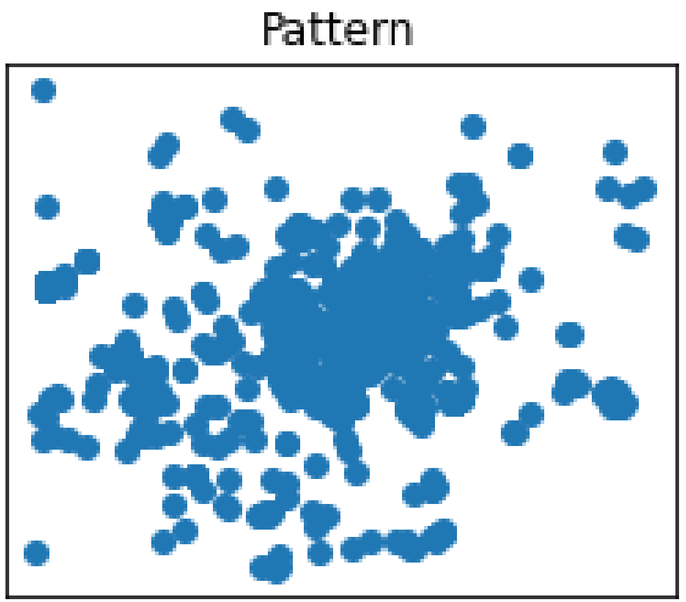
$G_{rand}(2) = 0/4 = 0$

$G_{rand}(3) = 4/4 = 1$

G function measures clustering for a distance *d*

Sample many times to get a simulation envelope with a confidence level





For those distances where the observed curve is above the envelope, we have significant clusters

Distance-based approaches – F function

- Generates a few random points ***P*** in the study area
- Is the fraction of random points which have a neighbor from the data within distance ***d***
- Measures empty space for a distance ***d***
- Determines the minimum distance from each random point in ***P*** to any original points ***O*** in the study area.

$$F(d) = \frac{sum[d_{min}(p_i, s) < d]}{n} \quad (13)$$

where *F(d)* indicates the value of the *F* function at distance *d*, and $sum[d_{min}(p_i s) < d]$ is the number of points in *P* with a minimum distance to any point in *O* smaller than *d*. The advantage of the *F* function is that one can increase the number of the randomly generated points to obtain a smoother curve.

F function measures empty space in distance *d*

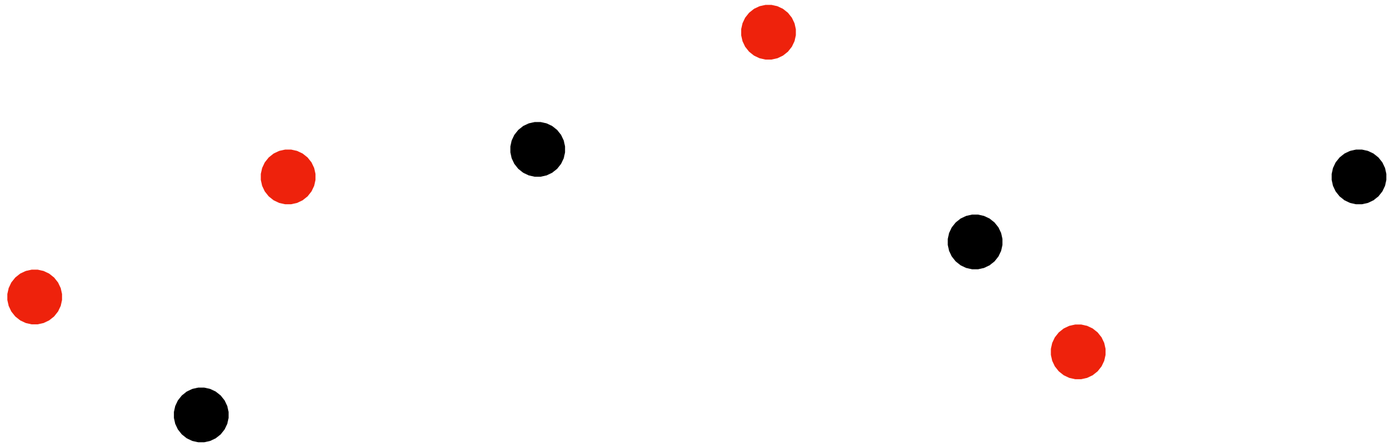
It is also called the "*empty space function*" because it measures the typical distance from arbitrary points in empty space.

It can be better to use F(d) than G(d) when there are few points.

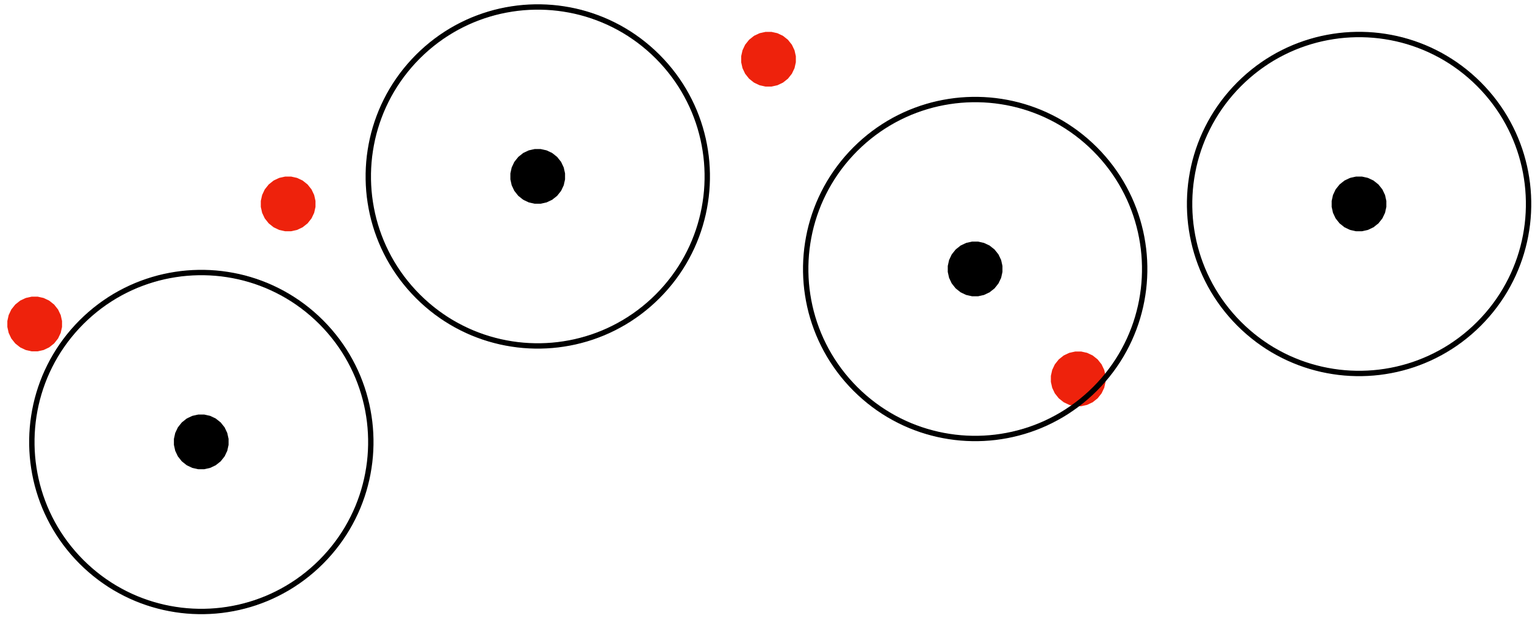
F function measures empty space in distance *d*

Data points

Random points

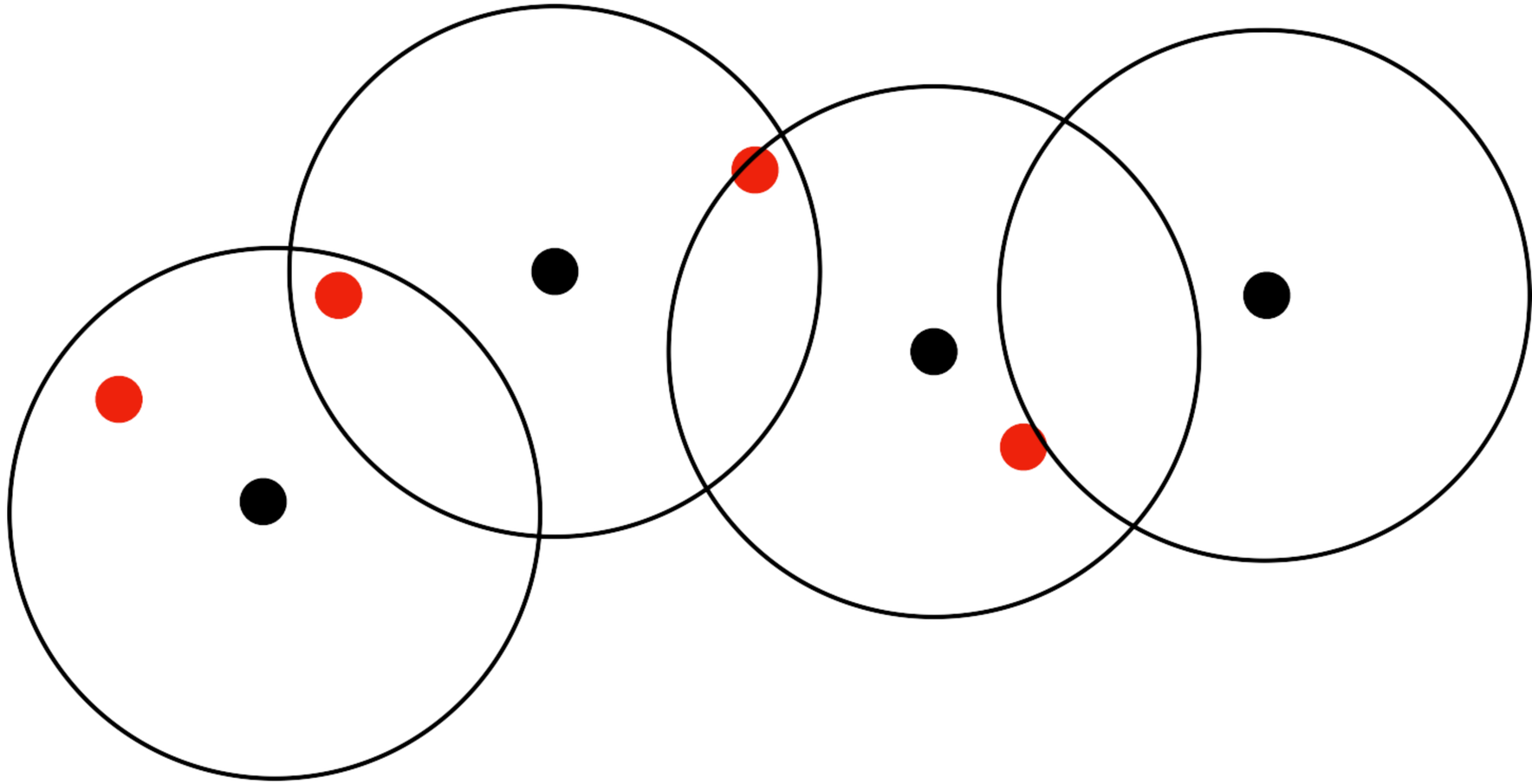


F function measures empty space in distance *d*



$$F(1) = 1/4 = 0.25$$

F function measures empty space in distance *d*

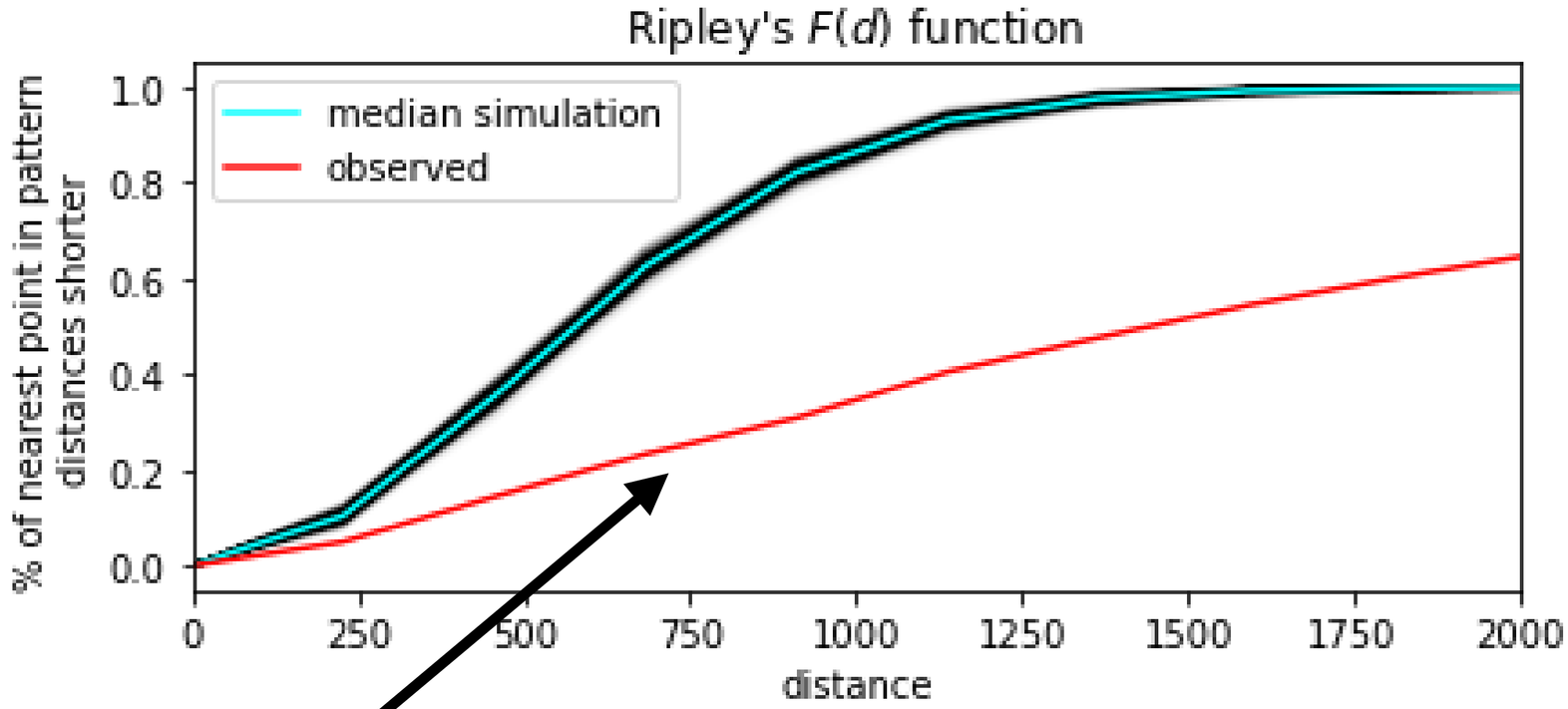


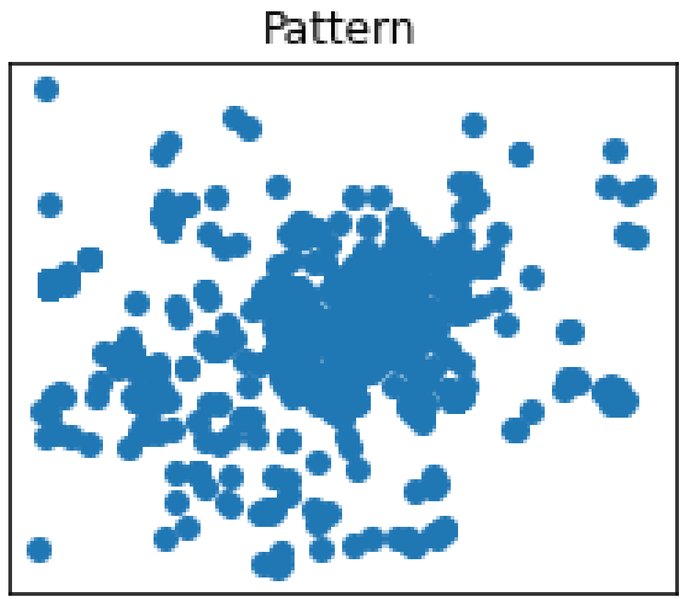
$F(1) = 1/4 = 0.25$

$F(2) = 4/4 =$

F function measures empty space in distance *d*

Sample many times to get a simulation envelope with a confidence level





For those distances where the observed curve is below the envelope, we have significant gaps

Distance-based approaches – F and G functions

F(d) and G(d) are first-order

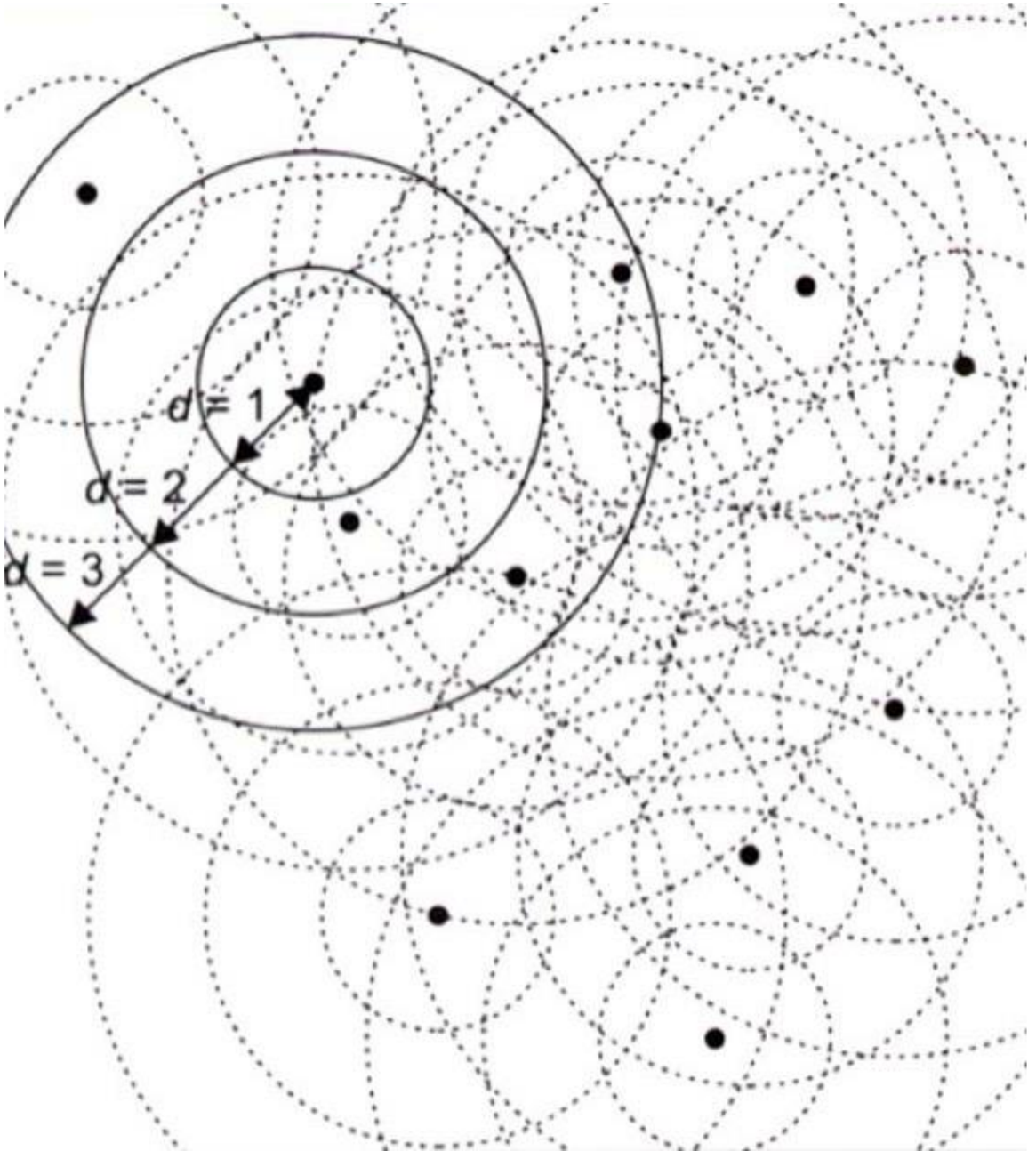
- First-order processes consider that all points happen as separate occurrences without interactions between positions

- To model second-order processes with interactions like competition (e.g. territory) or synergy (e.g. seed dispersal) use other measures like K(d).

Distance-based approaches – K function

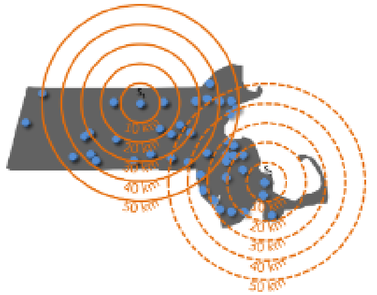
Alternative → K Function

- Construct a circle with a radius ***d*** around each point ***i***
- Count the total number ***n*** of points that fall inside any of the circles (excluding the points at the circle centers)
- Increment ***d*** by a small fixed amount and repeat the first two steps.



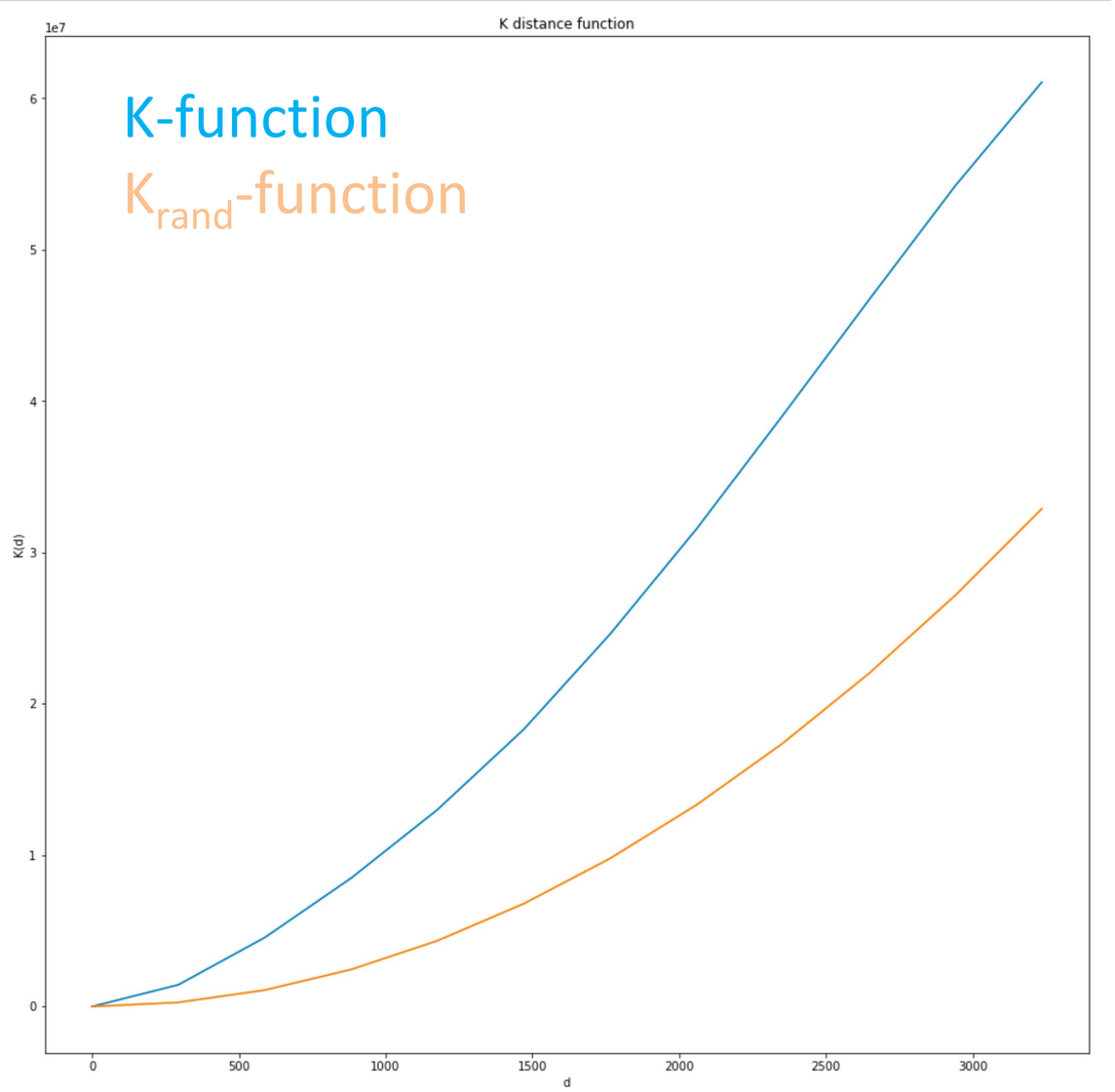
Distance-based approaches – K function

- Considers all distances between all events
- At each event, draw a circle of radius ***d*** and count the number of events in the circle
- Calculate the mean count for all locations ***i*** at that distance
- Divide mean count by the overall study area density to get ***K(d)***
- Repeat for all desired values of ***d***



Distance band (km)	# events from s_1	# events from s_2	# events from s_i	K
10	0	1	…	0.012
20	3	5	…	0.067
30	9	14	…	0.153
40	17	17	…	0.269
50	25	23	…	0.419

Distance-based approaches – K function



Interpretation

- If the line of the K-function for the pattern is above the CSR line, the pattern is clustered

- If it is below, it is dispersed

Modelling spatial point patterns

We can only observe the pattern.

How can we use this to find clusters?

- To infer the process we compare synthetic patterns from a null model with observed patterns. If different, we can reject the null model.

- Comparing the pattern to Complete Spatial Randomness (CSR) will tell us something about whether it is clustered or not

- CSR can generated by a homogeneous spatial Poisson

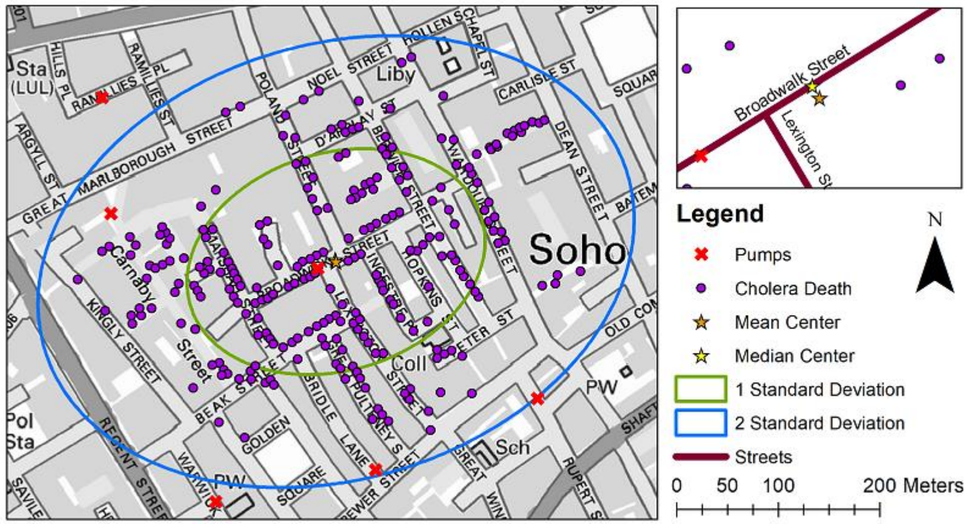
Modelling spatial point patterns

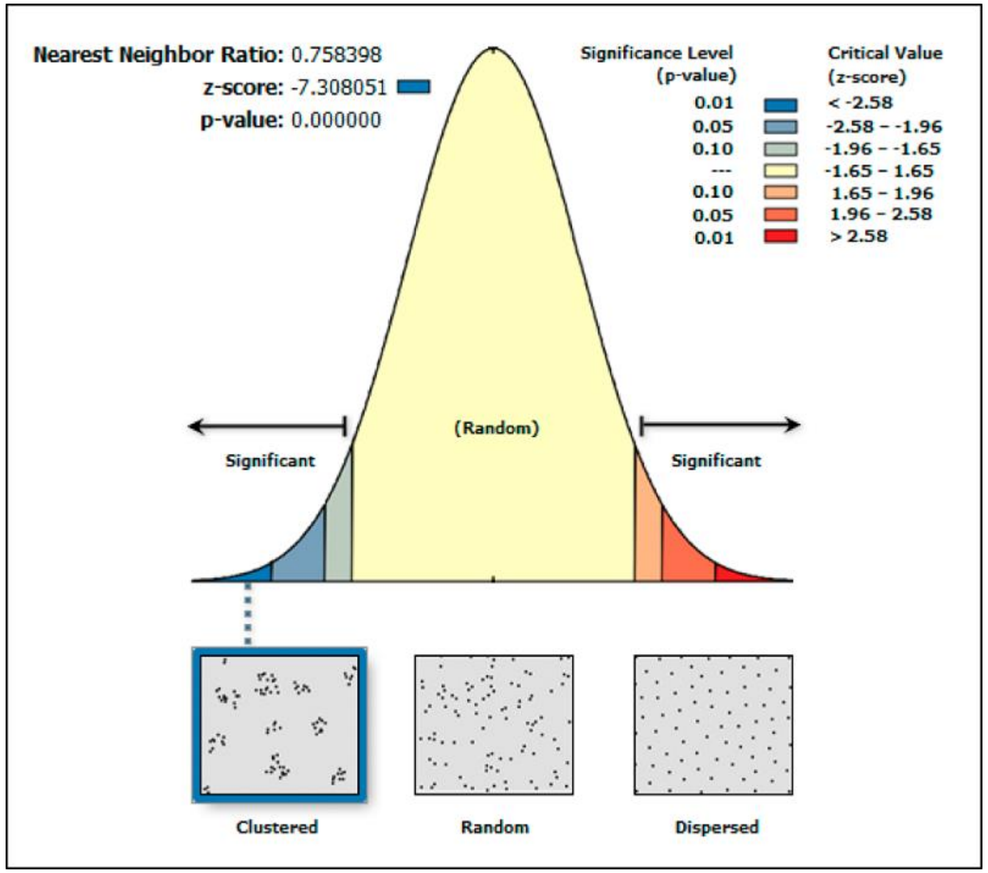
- A Poisson distribution is a discrete (countable) probability distribution.
- Poisson distribution models the probability of the number of events occurring over a fixed time step given a known mean.

$$P(k) = {\lambda^{\mathrm{k}} e^{-\lambda}}/{k!}$$

P – is the probability that an event will occur k times
k – is the number of events
λ – is the average intensity of the pattern
e – is the Euler's constant ≈ 2.718 (the base on the natural logarithmic system)

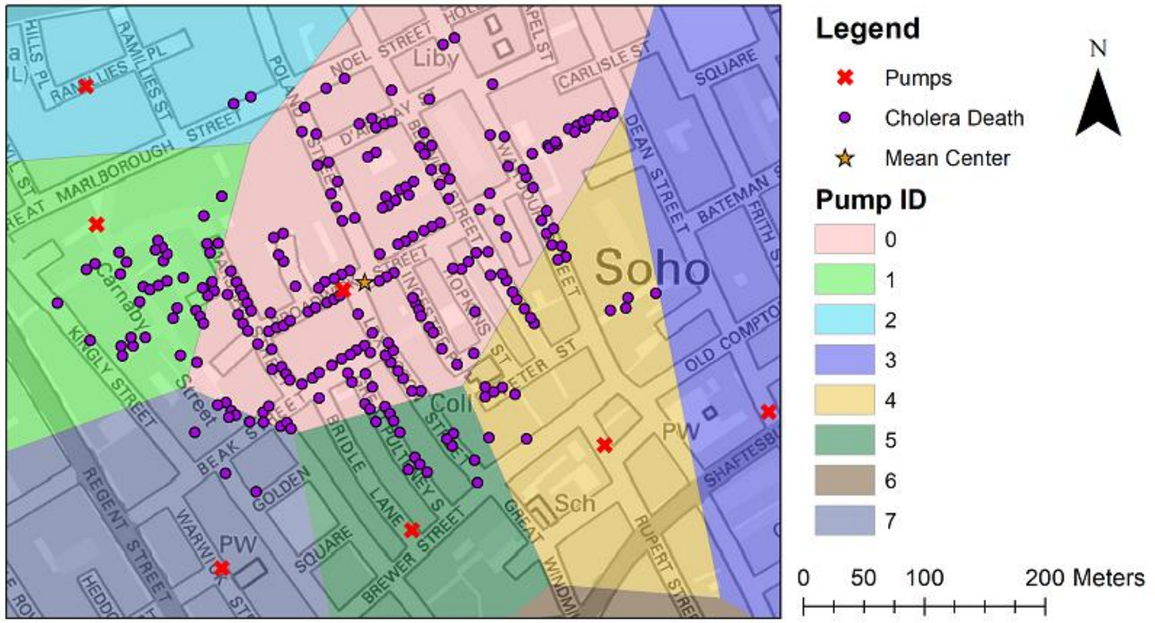
Revisiting the cholera outbreak in London

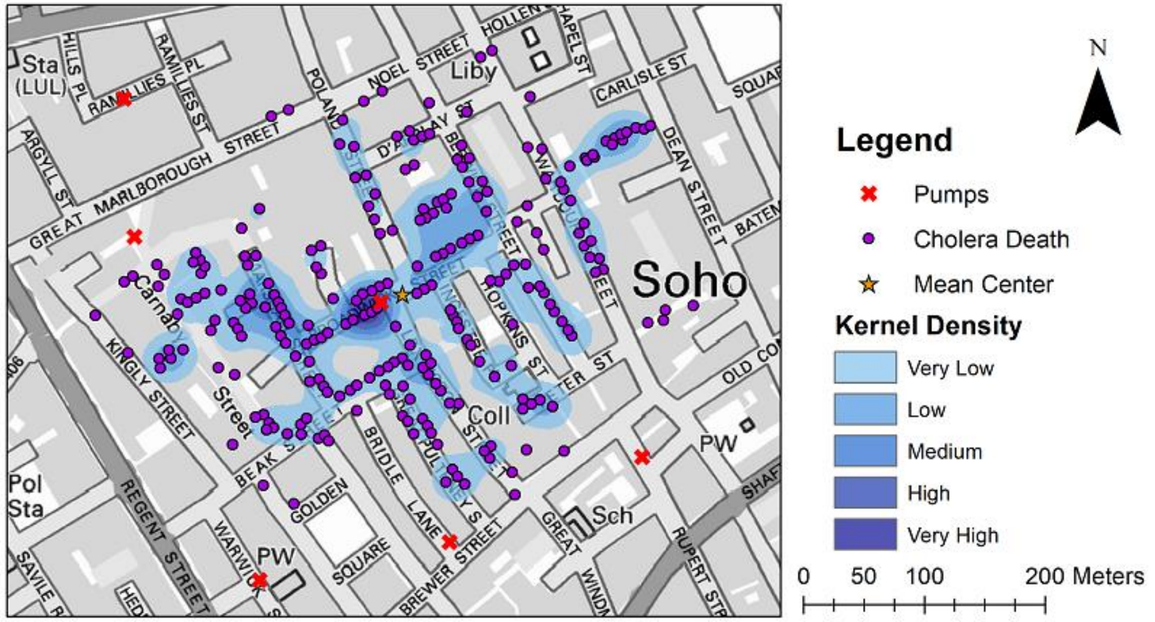




🔗 *https://gistbok.ucgis.org/bok-topics/point-pattern-analysis*

Revisiting the cholera outbreak in London



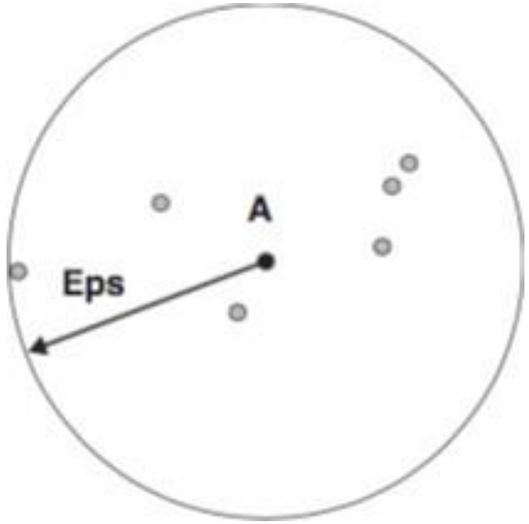


🔗 *https://gistbok.ucgis.org/bok-topics/point-pattern-analysis*

Finding clusters with DBSCAN

Density-based spatial clustering of applications with noise

DBSCAN: Density-based clustering



density of A = 7

1) Select a radius Eps

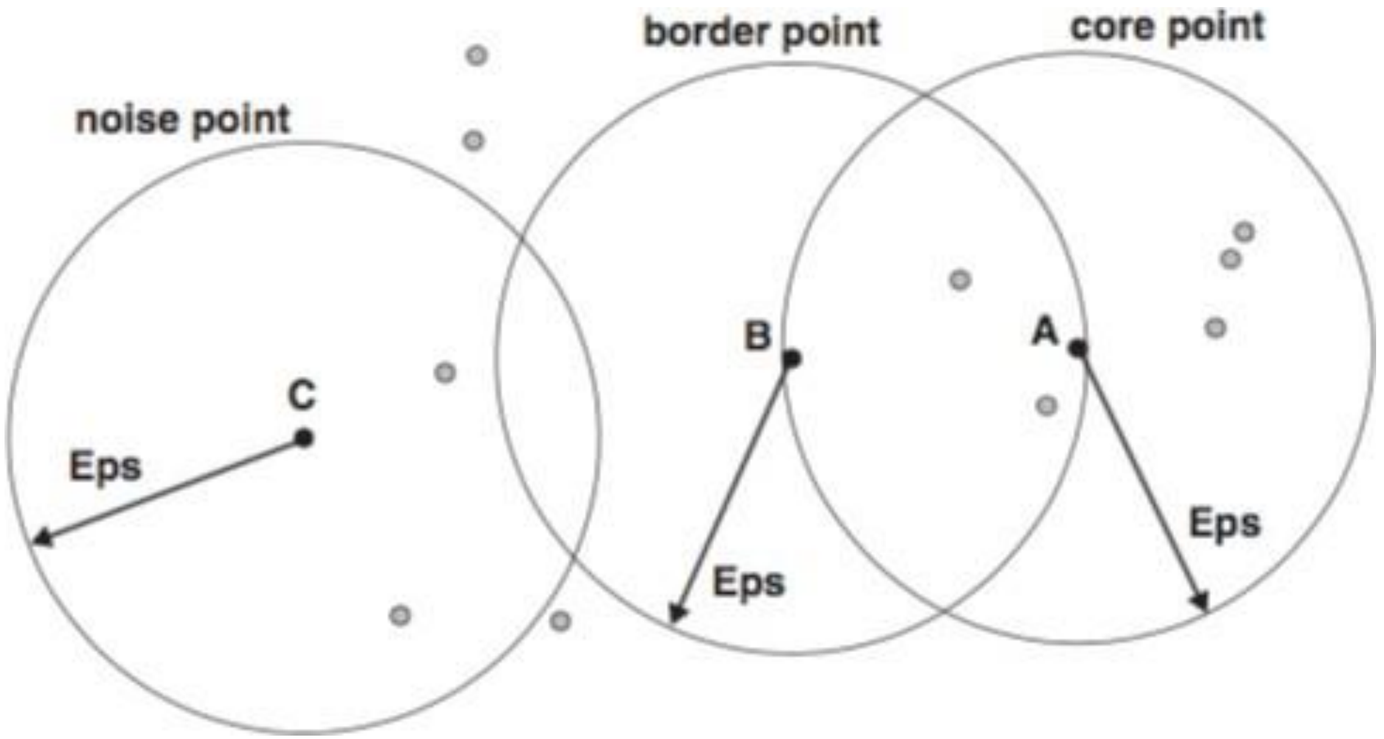
2) Select a parameter MinPts

3) Count all points within Eps (including the central point itself). This is the density of the point.

DBSCAN: Density-based clustering

1) Select a radius Eps
2) Select a parameter MinPts

3) Count all points within Eps (including the central point itself). This is the density of the point.



Core point: Has density ≥ MinPts

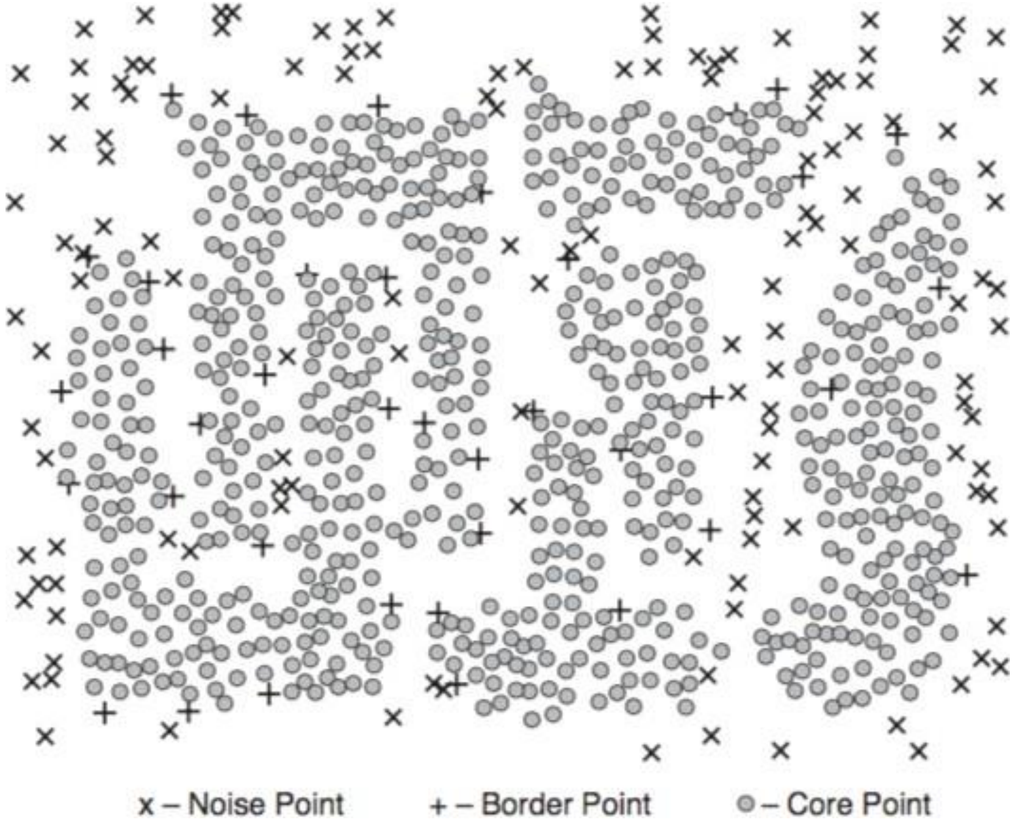
Border point: No core point but is in the neighborhood of a core point

Noise point: Neither core nor border point

DBSCAN: Density-based clustering

Algorithm 8.4 DBSCAN algorithm.

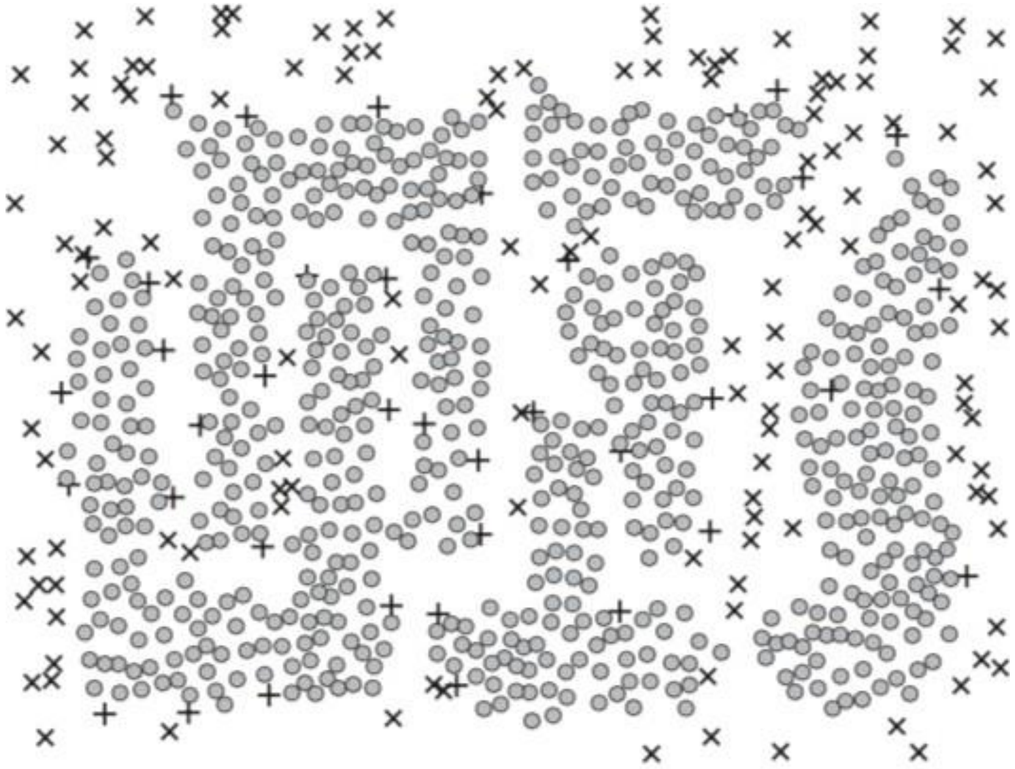
1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.



DBSCAN: Density-based clustering

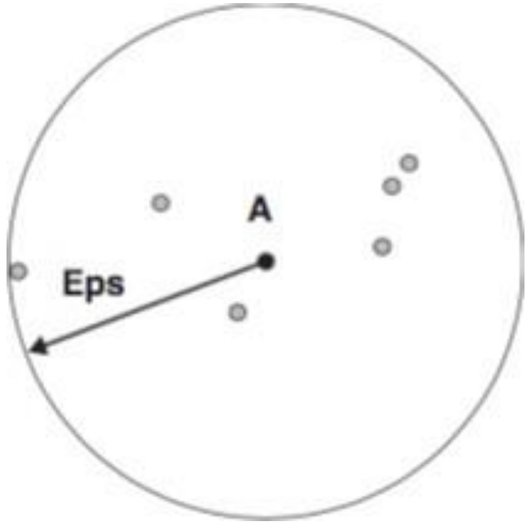
Algorithm 8.4 DBSCAN algorithm.

1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.



This is a partial clustering: Not all data points are assigned a cluster!

DBSCAN: Density-based clustering



density of A = 7

1) Select a radius Eps

2) Count all points within Eps (including the central point itself). This is the density of the point.

What are the extreme cases for the radius?

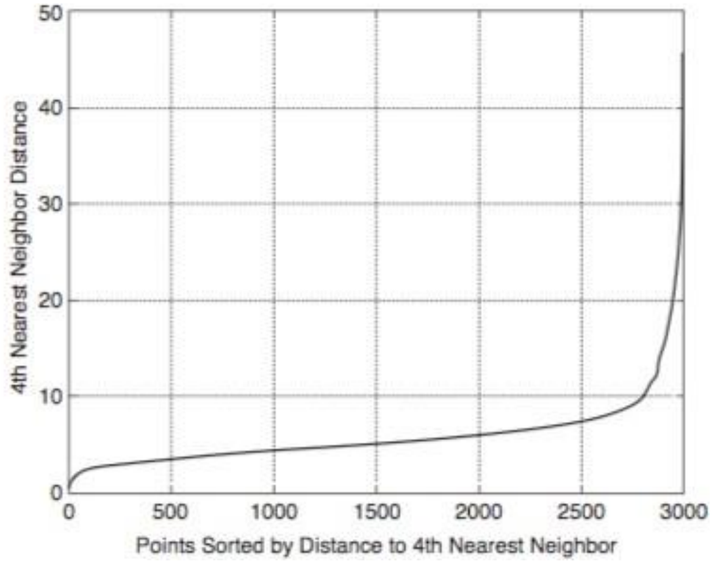
Too small: Every point has density 1

Too big: Every point has density n

DBSCAN: How to choose parameters?

How to choose Eps and MinPts?

k-dist plot k=4

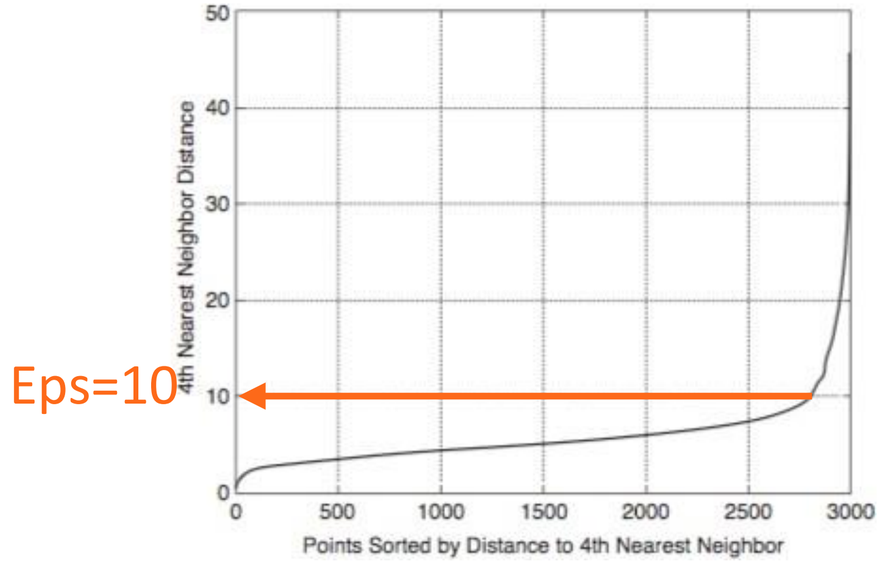


Look at the behavior of the distance from a point to it’s kth nearest neighbor

DBSCAN: How to choose parameters?

How to choose Eps and MinPts?

k-dist plot k=4

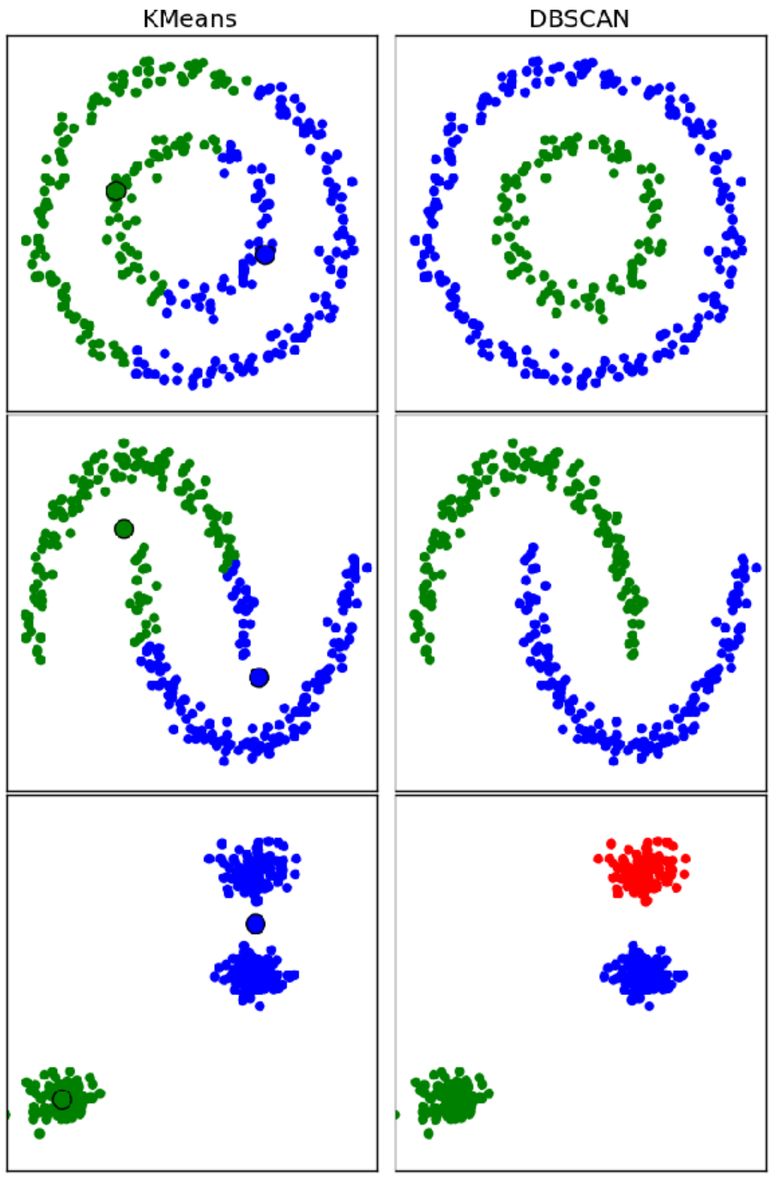


Look at the behavior of the distance from a point to it's kth nearest neighbor

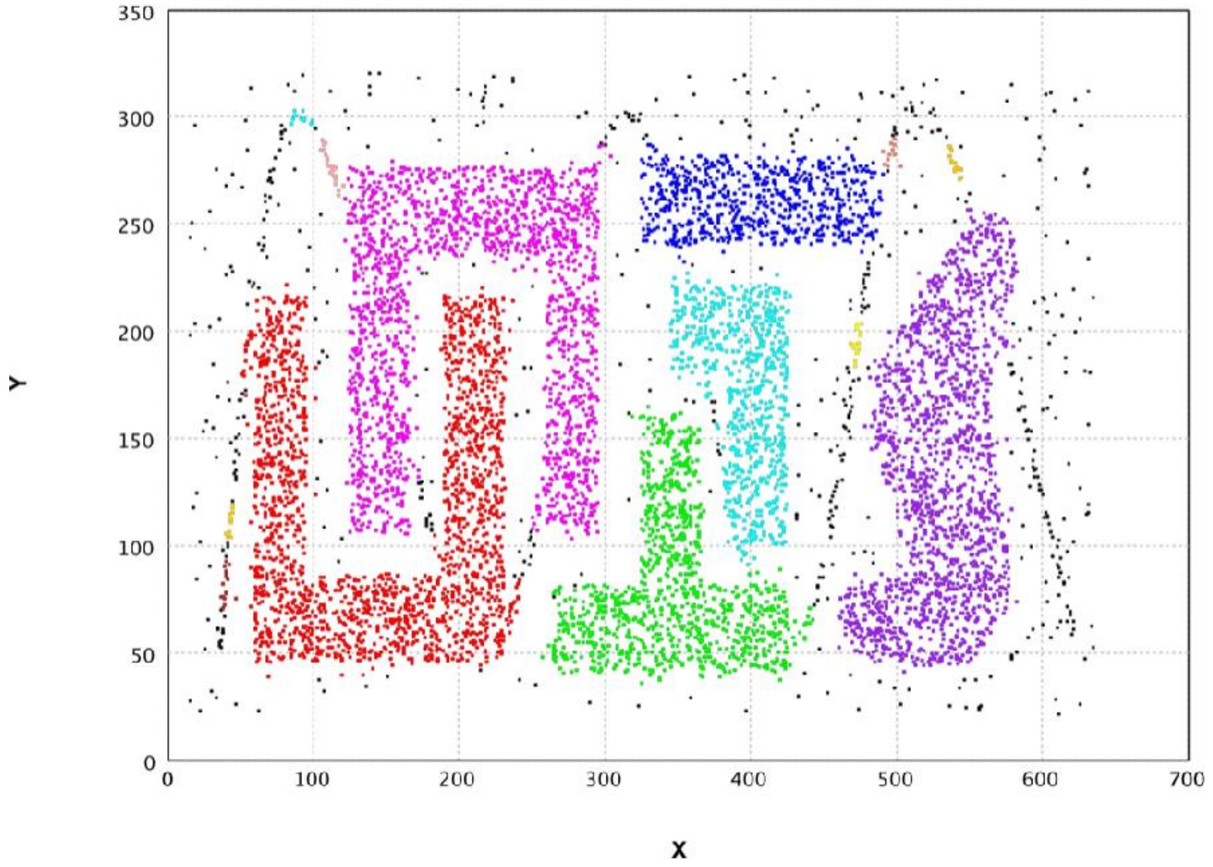
Noise points will have large distance. If they are systematic, there is a bend. This gives us a good Eps for the given k.

DBSCAN: Advantages

- Resistant to noise
- Can handle clusters of arbitrary shape and size
- Not necessarily spatial



DBSCAN: Disadvantages



- Cannot handle clusters with different densities
- Not based on a probabilistic model like LISA: no inference, no significance
- DBSCAN can be computationally expensive $O(n^2)$, but faster than most point pattern methods

Sources and further materials for today's class

Geographic Data Science with Python

https://geographicdata.science/book/ notebooks/08_point_pattern_analysis.html

https://darribas.org/gds_course/content/bH/concepts_H.html

Next week: OpenStreetMap

That’s all for today!!!

Next time together, field-based analysis (raster)
Until then, questions at LearnIT or geor@itu.dk