Cluster analysis

Mark Green

DASC507 – Advanced Biostatistics II

Analysis Methods for Complex Data Structures

Outline

- Dealing with point data
- Spatial weights
- Global Moran’s I
- Local Moran’s I
- Criticisms
- Case study

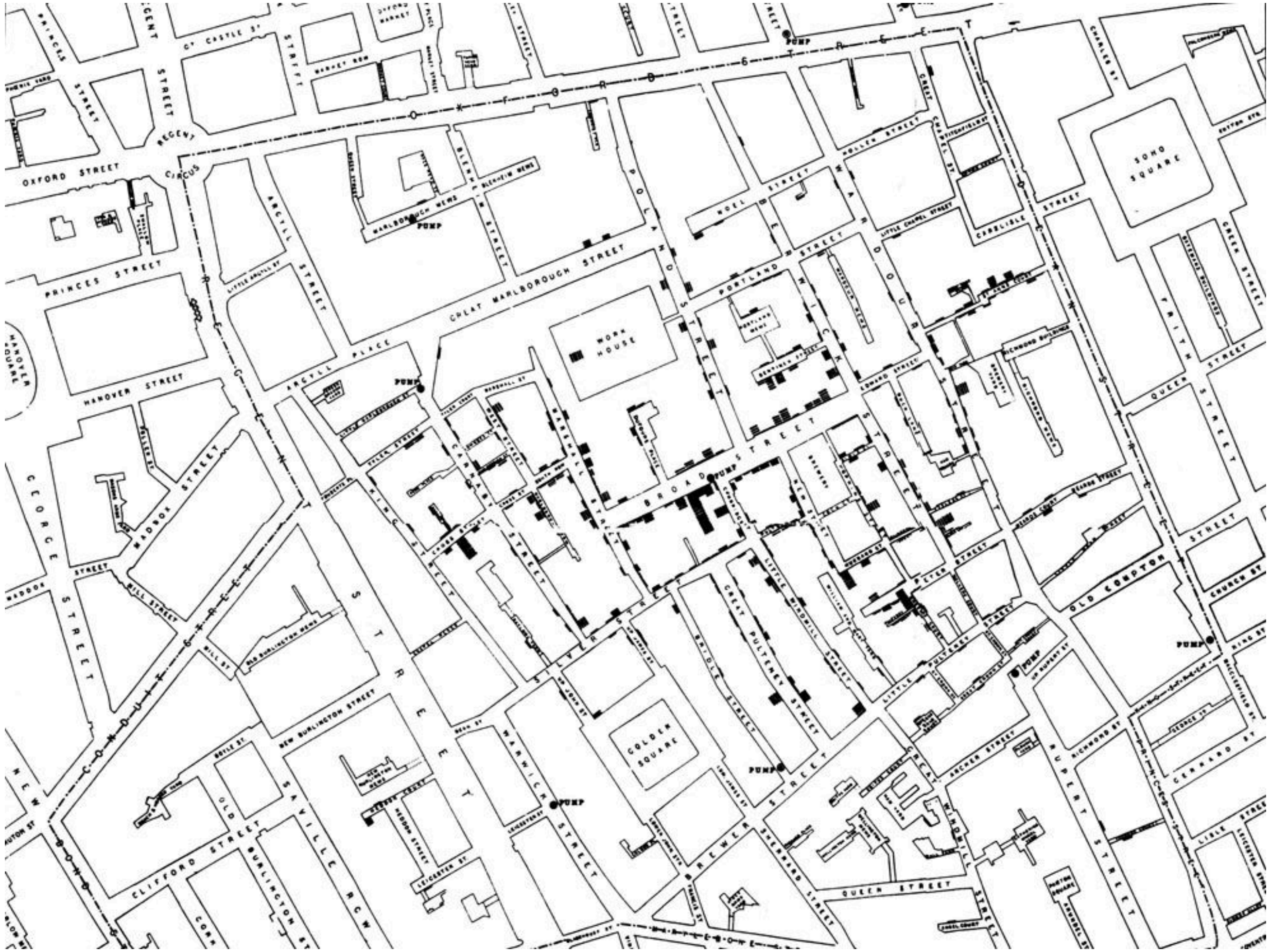
Introduction

What do we mean when we talk about 'clusters'?

- Dictionary definition "a group of similar things or people positioned or occurring closely together"
- Spatial proximity is explicit – concentration in places
- Need to identify if occurred by random chance (size can be key) or explained by other geographical phenomena
- How we can identify clusters will depend on the type of data…

Dealing with point data

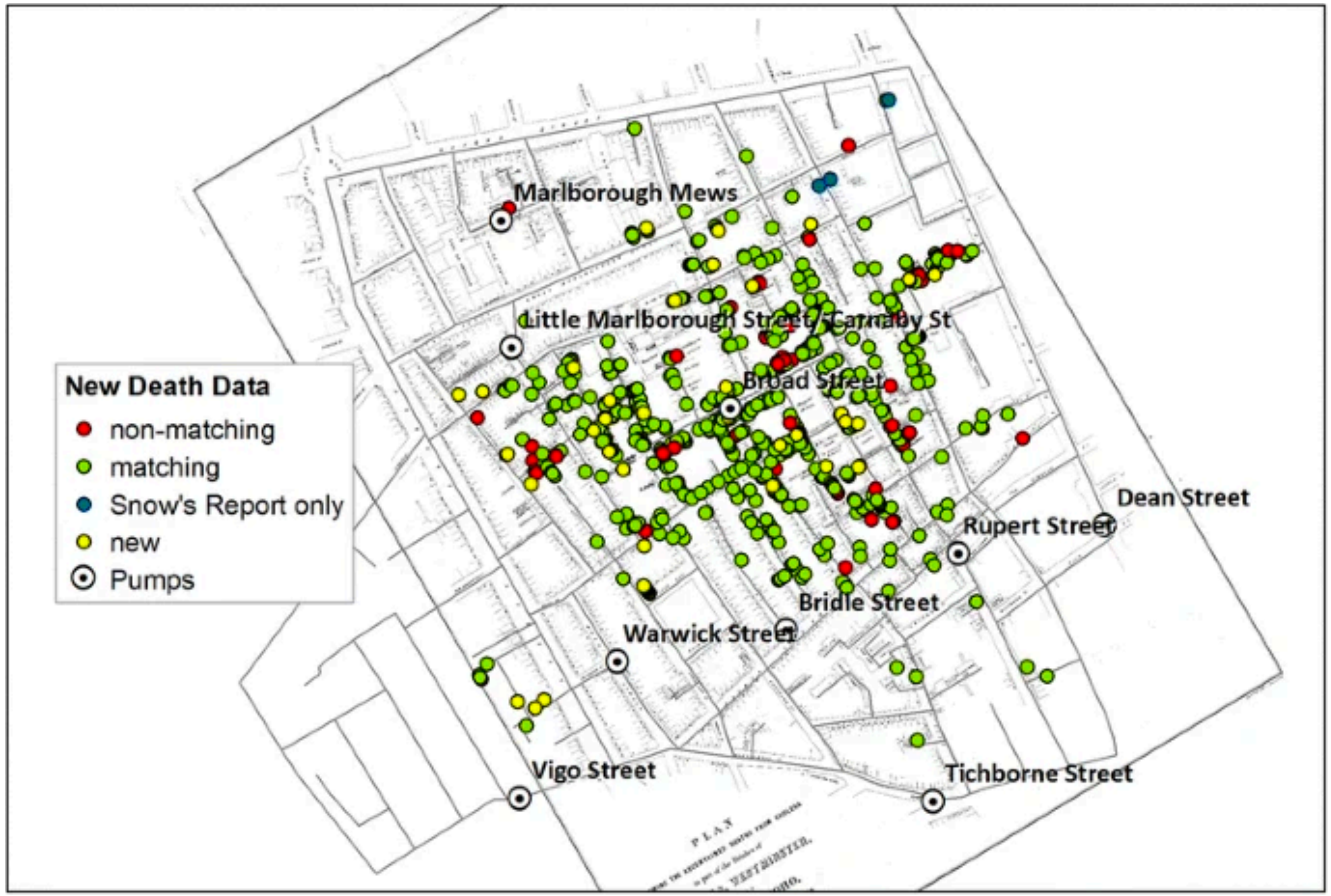
We can plot points, but how do we know if their locations are clustered?



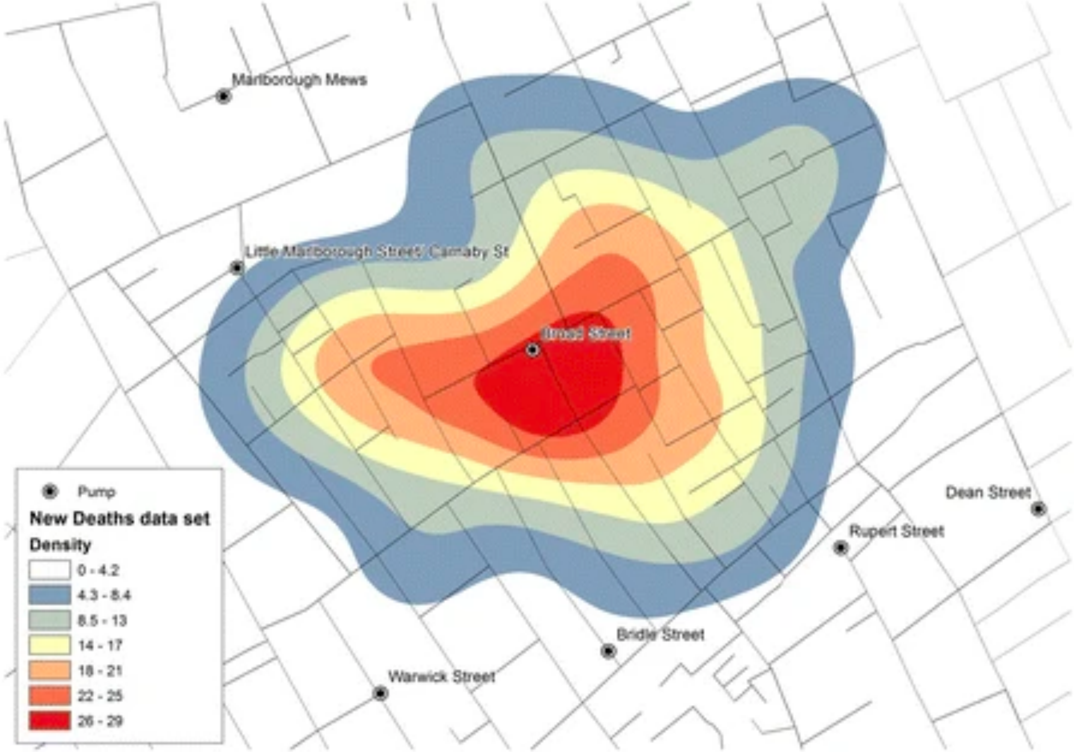
Dealing with point data

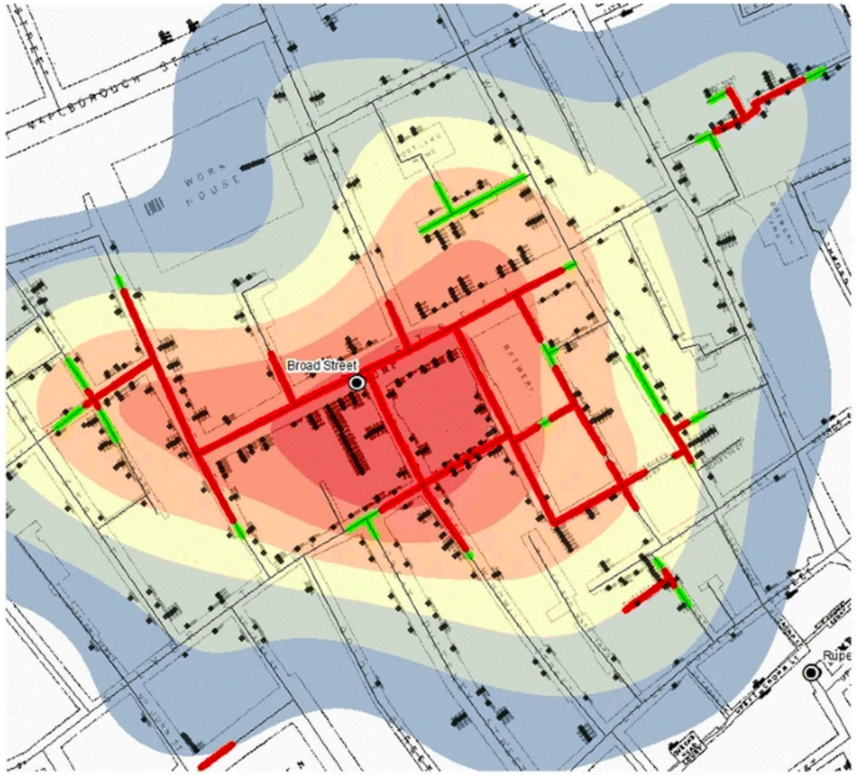
Let’s take an example…

Shiode et al. 2015. The mortality rates and the space-time patterns of John Snow’s cholera epidemic map. *International Journal of Health Geographics* **14**: 21.



Dealing with point data

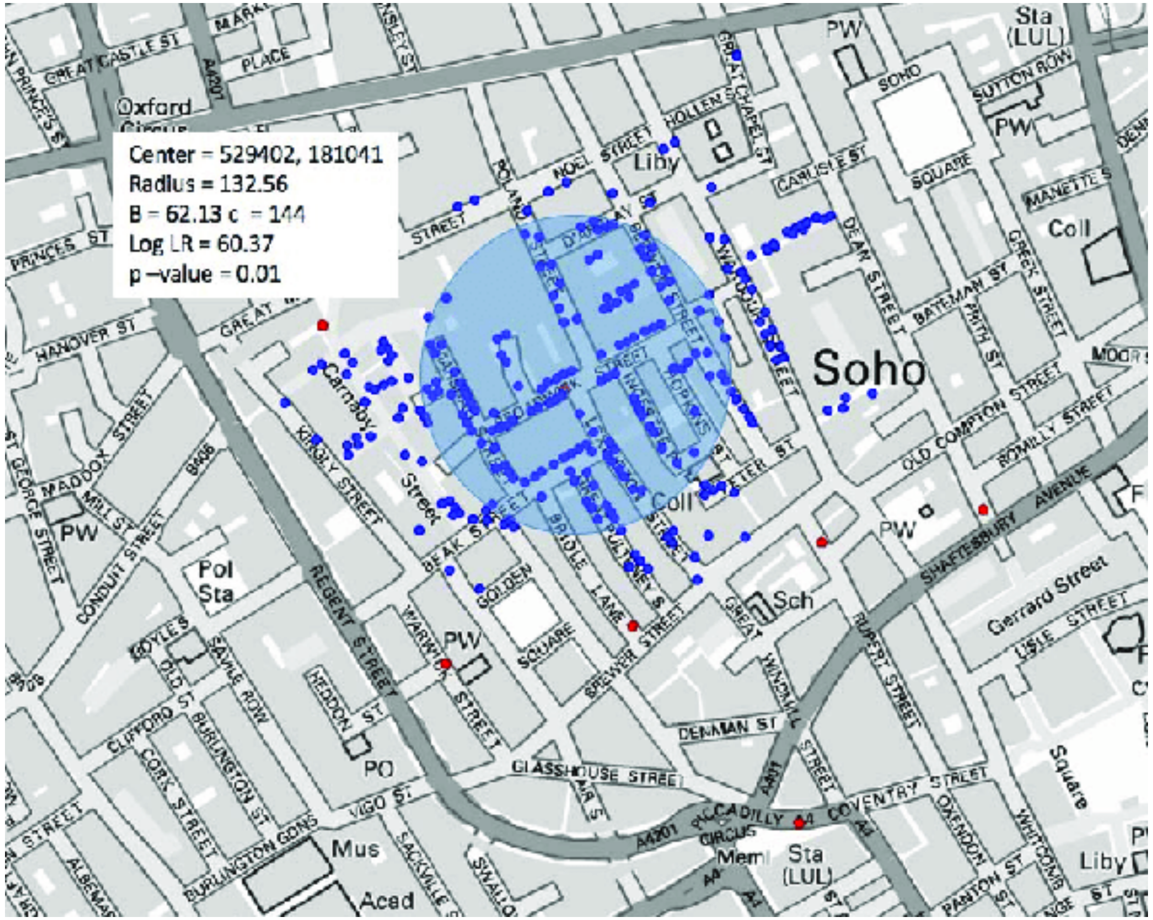




Kernel density estimation one approach for smoothing over space to generalise spatial patterns (and describe possible clusters)

Dealing with point data

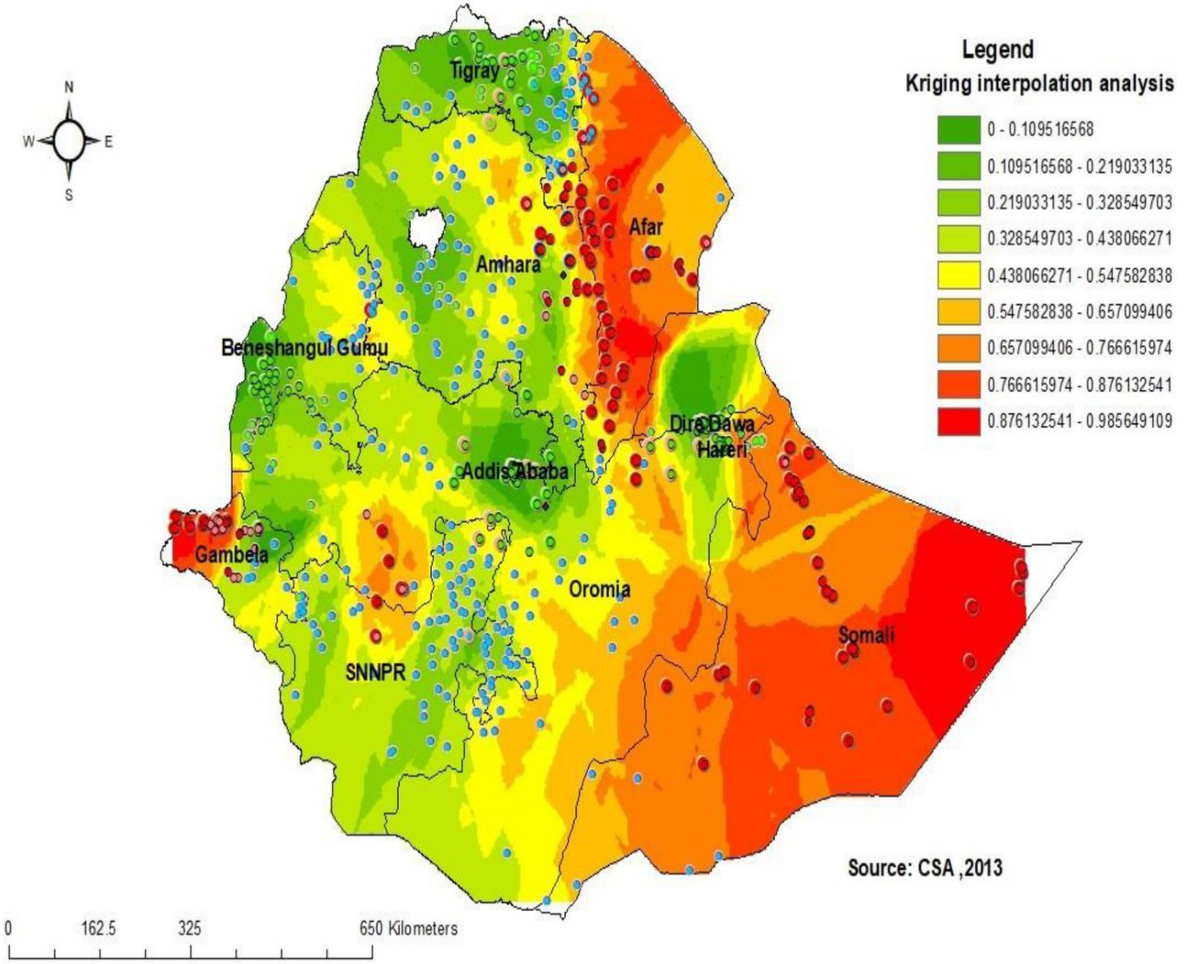
There are more formal ways of testing for spatial clustering of point data (e.g., SatScan)



Prasad et al. 2017.
https://doi.org/10.1109/BigDataCongress.2017.39

Dealing with point data

If we have sparse data points, we might try to estimate values inbetween using spatial interpolation methods – lots of methods for this



Agegnehu CD, Alem AZ. 2021. Exploring spatial variation in BCG vaccination among children 0–35 months in Ethiopia: spatial analysis of Ethiopian Demographic and Health Survey 2016. *BMJ Open* **11:**e043565. http://doi.org/10.1136/bmjopen-2020-043565

Spatial autocorrelation

- Data in one area are correlated to the data in areas surrounding them
- Each data point should have it’s own (unique) location
- Define which locations are next to each other
- How to measure distances between data points

Spatial weights

How might we represent the spatial structure of data if we want to model it statistically?

- Need to make spatial location explicit
- Each data point should have it’s own (unique) location
- Define which locations are next to each other
- How to measure distances between data points

Spatial weights

To formally define the spatial structure of data, we define W as

- N x N (positive) matrix for all data points
- Contains information over if each point i is related to each other point j
- Commonly, $W_{ij} > 0$ if i and j are neighbouring, else $W_{ij} = 0$ (not neighbours)
- W_{ii} is always 0
- How can we define a neighbour?

Spatial weights

Contiguity-based weights

- Which areas are ‘next’ to others
- Defined based on sharing a common boundary
  - Rook – must share long boundary
  - Queen – any point is touching
- If an area shares a common boundary, then Wij = 1
- *If not a neighbour, then Wij = 0*

Spatial weights

Distance-based weights

- How close are other areas
- W defined as (inversely) proportional to distance between each data point
  - Distance (e.g., how near are each area) – may have threshold cut-point
  - Distance as buffer (e.g., select areas within 500m of an area)
  - K-Nearest Neighbours (e.g., select *k=4* nearest areas only)
- Wij can be either binary (e.g., Wij = 1 if defined by KNN) or as continuous if based on distance (0 < Wij < 1) where larger values are closer
- If not a neighbour, then Wij = 0

Spatial weights

Block weights

- Is the area within the same ‘place’ as other areas
- Weights depend on whether areas are located in particular (larger) areas
  - Small administrative zones located within larger zones
  - Postcode within a electoral ward/city
- If an area is located within a particular zone, then Wij = 1
- *If not a neighbour, then Wij = 0*

Spatial weights

Selecting the correct weight

- Depends on the nature of the spatial process you are measuring
- If process has an immediate impact, then contiguity will work best
- If need to incorporate a wider spatial extent (e.g., accessibility) then distance-based weights preferred
- If data are multi-level or reflect a specific process by area structure (e.g., policies implemented in certain areas) then use block-based
- If just interested in the statistical nature, contiguity often used

Spatial weights

Spatial weights may need to be standardized, especially if being modelled

- Most commonly used approach is row-based standardization

$$\overline{w_{ij}} = \frac{w_{ij}}{\sum w_i}$$

Spatial autocorrelation

- Data in one area are correlated to the data in areas surrounding them
- Each data point should have it's own (unique) location
- Define which locations are next to each other
- How to measure distances between data points

Global Moran's I

- Summary measure describing the overall nature of clustering in a dataset (i.e., describing the general spatial patterning of values)
- Visualised using a scatter plot (also termed Moran Plot)
  - X-axis: Spatially lagged outcome variable (y) – for each area I, calculate an average of y for surrounding areas by weighting values by W.
  - Y-axis: outcome variable value for an area
- This allows us to compare how similar an area (Yi) is to its surrounding neighbouring areas (WYi)
- Variable and spatial weights should be standardised

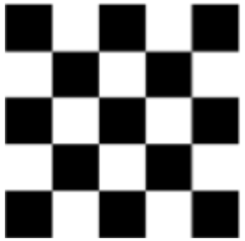
Global Moran’s I

We can calculate a single statistic to describe the presence of clustering (equivalent to the line of best for the Moran’s Plot).

$$I = \frac{N}{\Sigma_i \Sigma_j w_{ji}} \frac{\Sigma_i \Sigma_j w_{ij}(xi\ - xbar)(xj\ - xbar)}{\Sigma_i (xi\ - xbar)^2}$$

Values run from -1 to +1 where:

- +1 is perfect clustering (i.e., high values located by high values, *vice versa*)
- 0 is a random pattern (no clustering)
- -1 is a dispersed pattern of alternating high then low (rare) ->



Local Moran's I

- If we find existence of clustering overall, then the next question is *where* is that clustering located
- Clusters can be defined as:
  - High-high – hotspots or areas with spatial autocorrelation of high values (i.e., areas of high values, surrounded by areas of high values)
  - Low-low – coldspots or areas with spatial autocorrelation for lower values (i.e., areas of low values, surrounded by areas of low values)
  - High-low – a spatial outlier with an area of high value surrounded by lower values
  - Low-high – as above, but other way round

Local Moran’s I

- If we find existence of clustering overall, then the next question is *where* is that clustering located

$$I_i = \frac{(x_i - \bar{x})}{m} \sum_j w_{J1} \cdot \left(j_J^c - \bar{x}\right)$$

$$m = \frac{\Sigma_i (x_i - \bar{x})^2}{N}$$

- Monte Carlo approach for significance – compares observed patterns, with many simulated random patterns to see how likely it would be to find these observed patterns

Criticisms

- Formal definitions of spatial structure are rigid
  - Area size may distort patterns (e.g., a large area sharing a common boundaries has same weighting as a smaller one)
  - Spatial structure harder to formalize when measuring ‘communities’
- Descriptive tools that cannot help us to explain clustering processes (but knowing ‘where’ allows us to ask ‘why’)
- Neutral global Moran’s I can often hide local patterns
- Moran’s I statistics require continuous measures

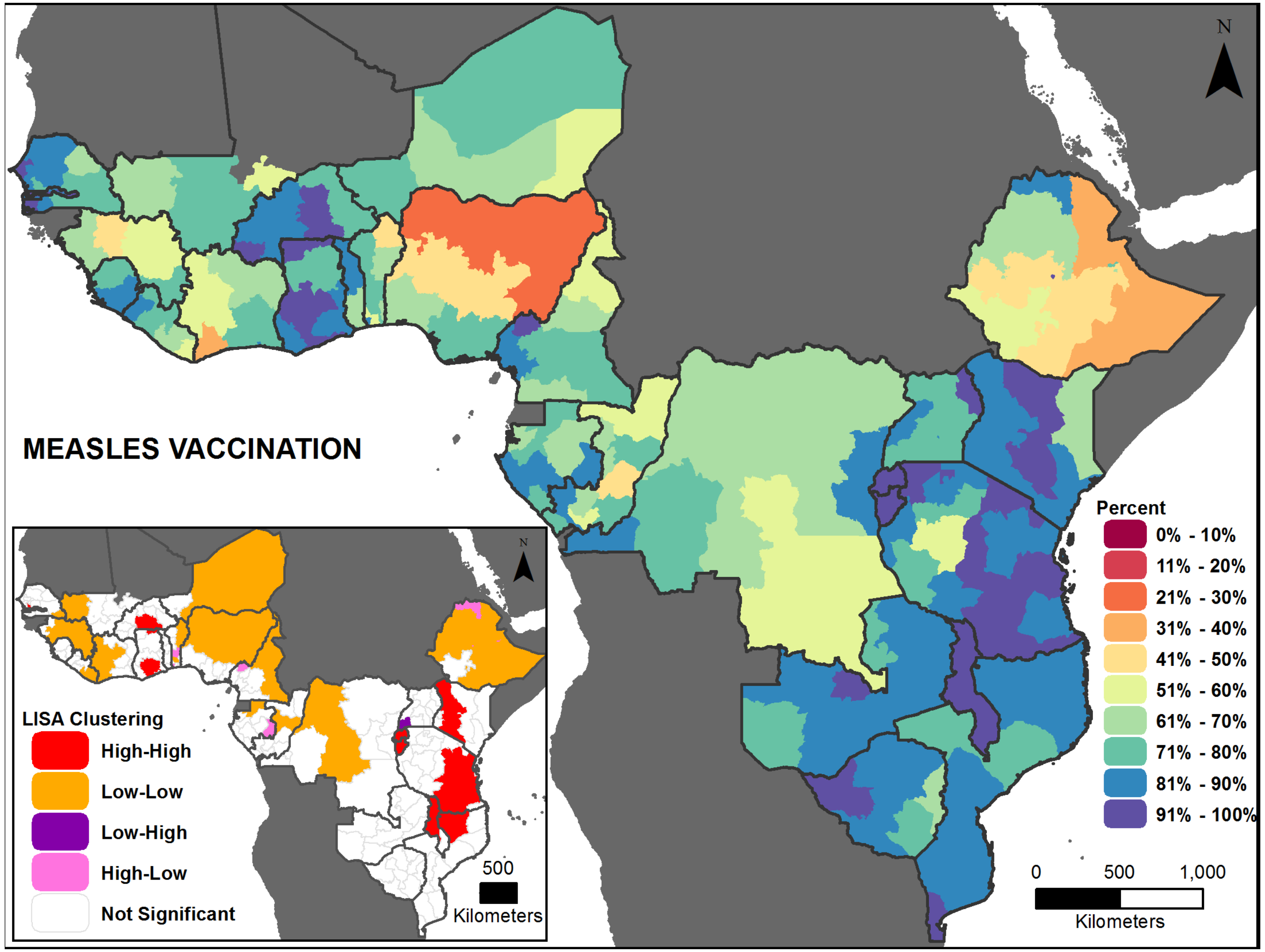
Case study

Yourkavitch et al. 2018. Using geographical analysis to identify child health inequality in sub-Saharan Africa. *PLoS One* **13**(8): e0201870.

- Mapping Sustainable Development Goals (SDG) important for tracking progress
- Aggregate measures for countries ignores spatial heterogeneity (including where progress may be further behind)
- Consider 6 SDG metrics for child health
- Sub-regional metrics estimated from Demographic and Health Surveys

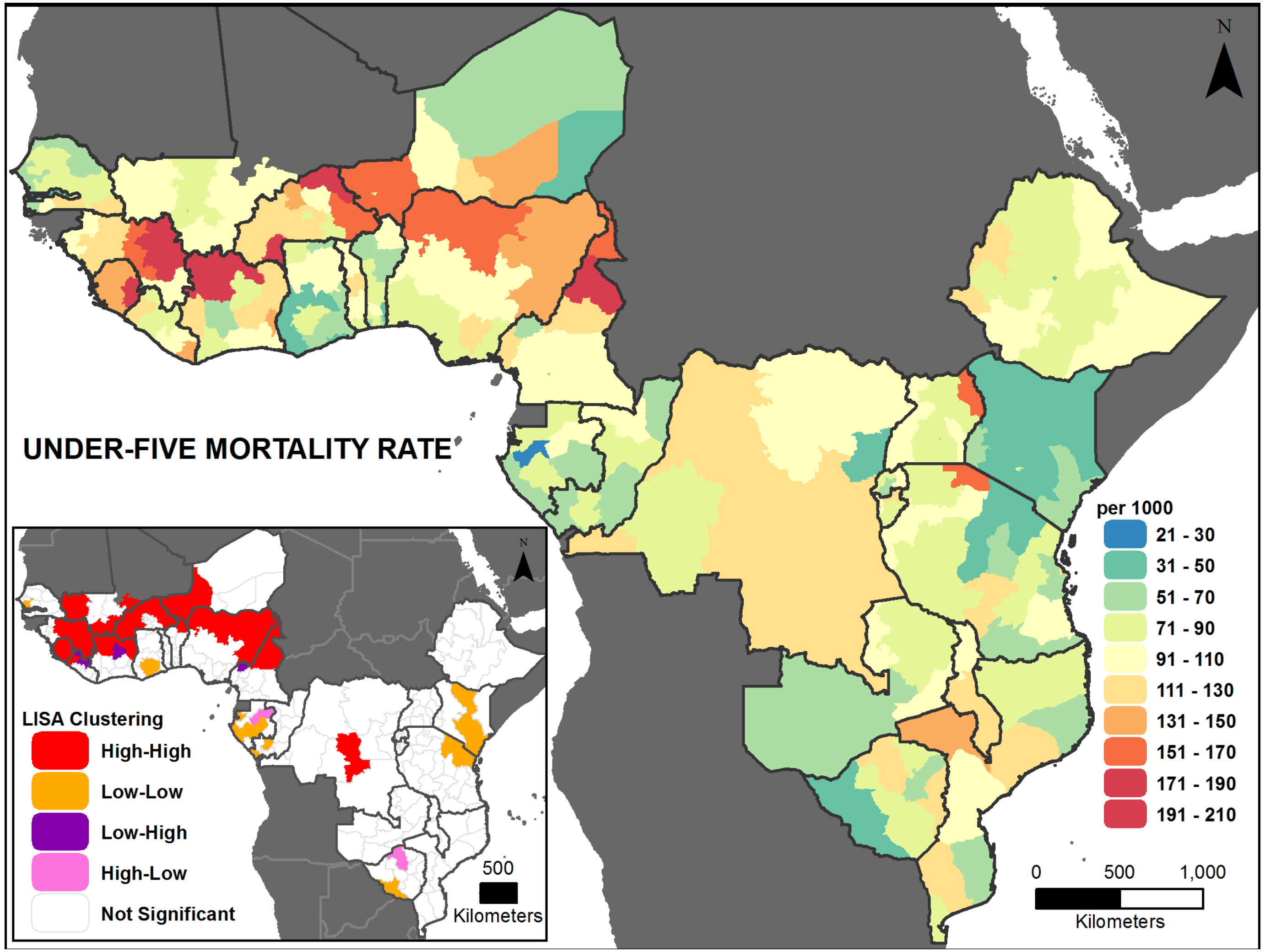
Case study

- DPT3 immunisations (measles, diptheria, pertussis and tetanus)
- I = 0.52
- Low coverage often in rural regions, but large heterogeneity across countries



Case study

- Under-5 mortality
- I = 0.41
- Higher rates in Western Africa, especially rural regions
- Also not distinct spatial pattern, suggests caution of focusing only on the statistic



Further reading

- Anselin L. 1995. Local Indicators of Spatial Association – LISA. *Geographical Analysis* **27**(2): 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x
- Carlos et al. 2010. Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics* **9**: 39. https://doi.org/10.1186/1476-072X-9-39
- Shiode et al. 2015. The mortality rates and the space-time patterns of John Snow's cholera epidemic map. *International Journal of Health Geographics* **14**: 21. https://doi.org/10.1186/s12942-015-0011-y
- Walther G. 2010. Optimal and fast detection of spatial clusters with scan statistics. Annals of Statistics **38**(2):1010-1033. https://doi.org/10.1214/09-AOS732