Spatial analysis of public health data

21/02/2022 (updated: 02/02/2024)

Andy MacLachlan

a.maclachlan@ucl.ac.uk

andymaclachlan

andrewmaclachlan

PDF presentation

Centre for Advanced Spatial Analysis, UCL

Slide acknowledgement: Professor Adam Dennett

How to use the lectures

- Slides are made with xaringan
- 🔍 In the bottom left there is a search tool which will search all content of presentation
- Control + F will also search
- Press enter to move to the next result
- ✏ In the top right let's you draw on the slides, although these aren't saved.
- Pressing the letter o (for overview) will allow you to see an overview of the whole presentation and go to a slide
- Alternatively just typing the slide number e.g. 10 on the website will take you to that slide
- Pressing alt+F will fit the slide to the screen, this is useful if you have resized the window and have another open - side by side.

Slide and content acknowledgement: Professor Adam Dennett

Outline

- The importance of patterns
- Patterns of categorical point data – Point Pattern Analysis
    - Quadrat Analysis
    - Ripley's K
    - DBSCAN
    - HDBSCAN
- Patterns of spatially referenced continuous observations
    - Spatial autocorrelation
    - Defining near and distant things
    - Measuring spatial autocorrelation
    - Moran's I
    - LISA (Local indicators of spatial association)

Broad street pump and John Snow pub. Source:Hartford Courant

Part 1: Point Pattern Analysis

Questions we can ask / set

Points

Are these points distributed in a random way or is there some sort of pattern (uniform or clustered)?

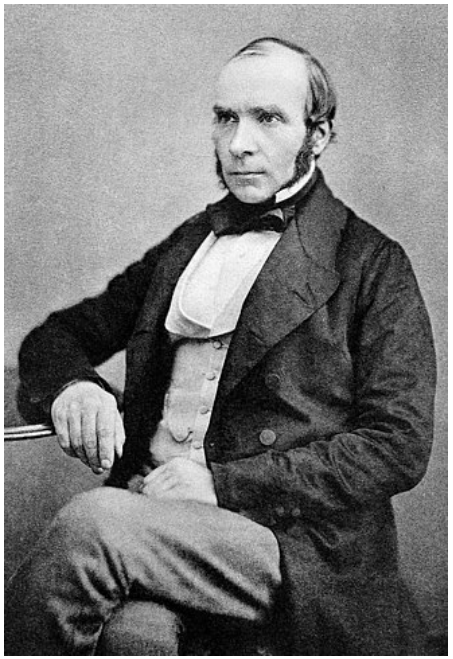
Spatially continuous observations (e.g. values of polygons)

How (dis)similar are our values assigned to geographic units across geographic space

The first (?) point pattern analysis

Dr John Snow

Jon Snow. Source: Wikipedia

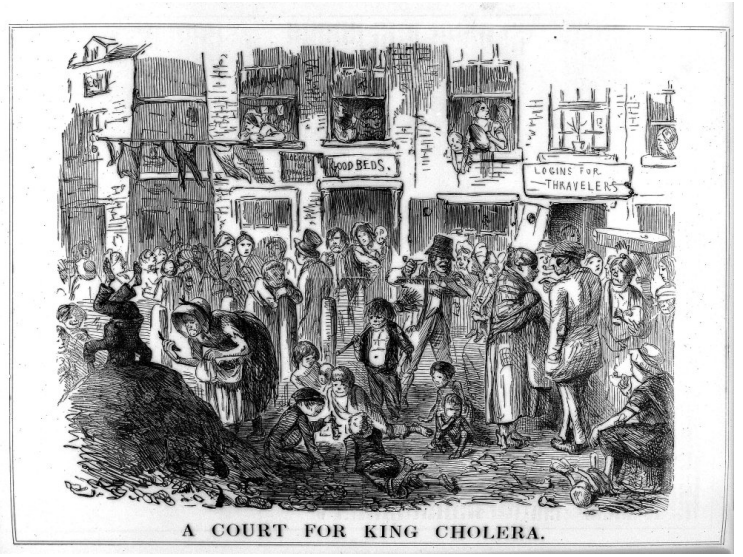


John Snow 1813-1858. Source:Wikipedia

Background

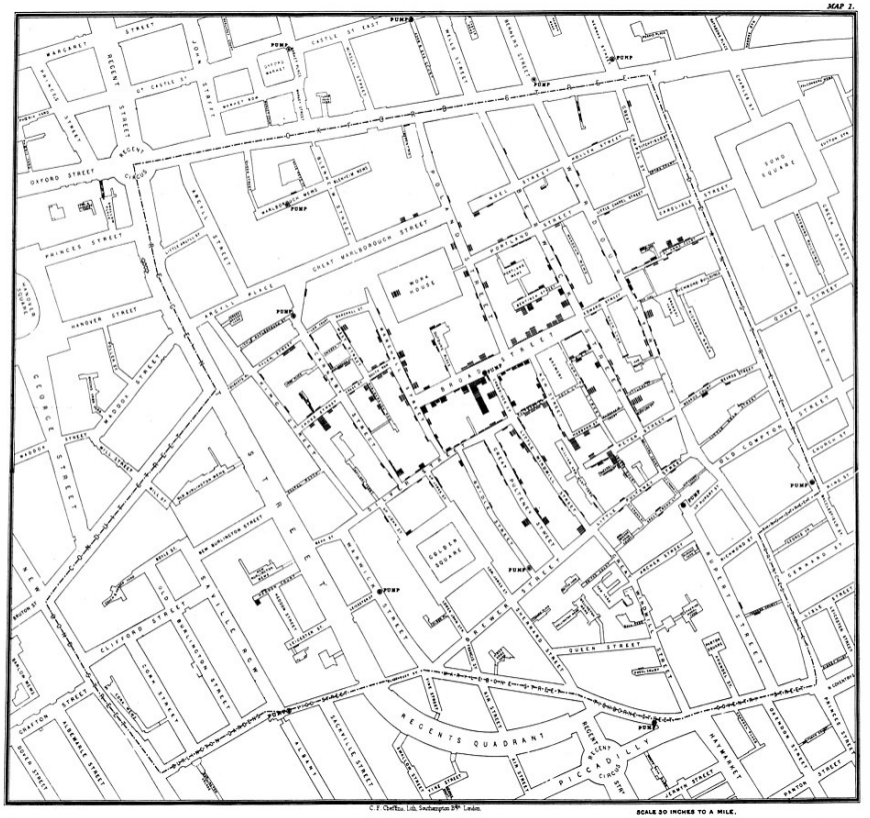
- In response to the Cholera outbreak in 1854
- This was the third outbreak the city had seen....
- John Snow saw a pattern from those dying of the disease ...
- a **spatial pattern**
- He was a doctor and pioneered the use of anesthesia

Cartoon (dated 1852) showing that cholera was from social overcrowding associated with the industrial revolution as opposed to polluted water...



A court for king cholera. Source:Science museum

Famous map!



Something in the water: the mythology of Snow's map of cholera. Source:Kenneth Field

History of point pattern

- John snow often attributed with identifying the outbreak at the broad street pump
- Data was from mortality reports **issued by the Registrar General Office not surveys**
- Snow wasn't responsible directly, but it contributed to the knowledge
- Snow also didn't invent this mapping style
- Snow didn't make the map, drawing was done by **Charles Cheffins** and probably others
- Valentine Seamen (1770-1817) did **before Snow was even born!**

Valentine Seamen ca 1800. Source:Brian Altonen

Valentine Seamen wondering if his work would be covered would be covered in a GIS class 220 years later...

The problem...

Benjamin Rush

- We have Black Plague / Yellow Fever in Philadelphia (1793)
- Benjamin Rush (Founding father of USA) said it must be from foreign goods
- Believed it was contagious like COVID!

Philadelphia Under Siege: The Yellow Fever of 1793. Source: Samuel A. Gum

The problem...

New York Doctors

- Ships were the cause!
- But it wasn't passing from person to person
    - Those who cared for these people didn't get it!
- The ships must be the cause but the people on them have developed less contagious version!
- There was another source that was from the decaying material in the ships and dock area
- Once someone was dying this also spread the disease

Arch Street Wharf in Philadelphia, where some of the first cases were identified. Source: History Channel

Rush then claimed in was from a bad batch of coffee!

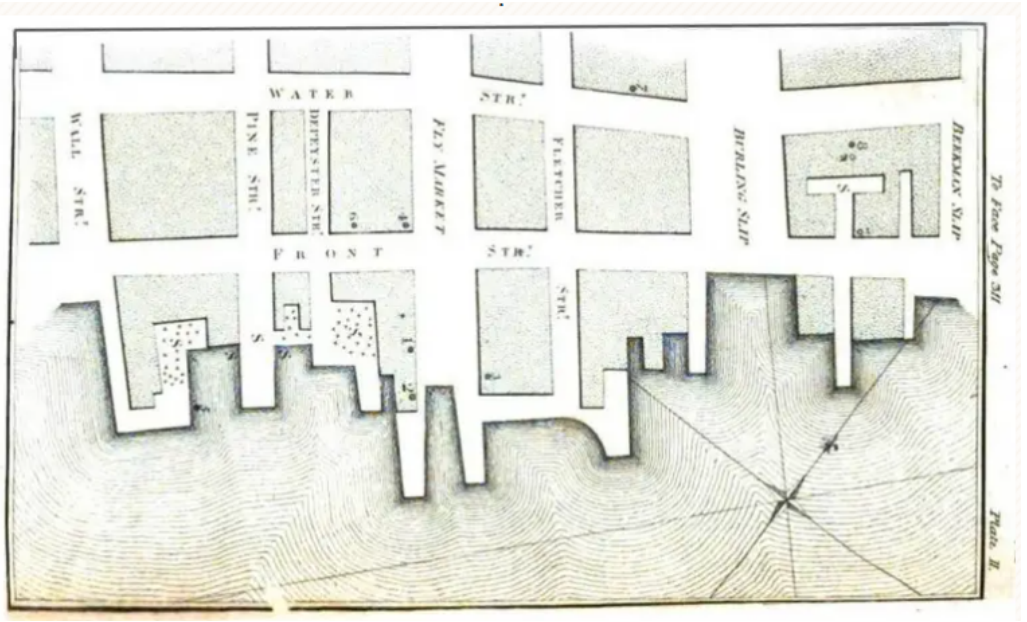
Ok Rush?....

The problem...

It is linked to ships, people, hot temperature, putrid air + water and cities!

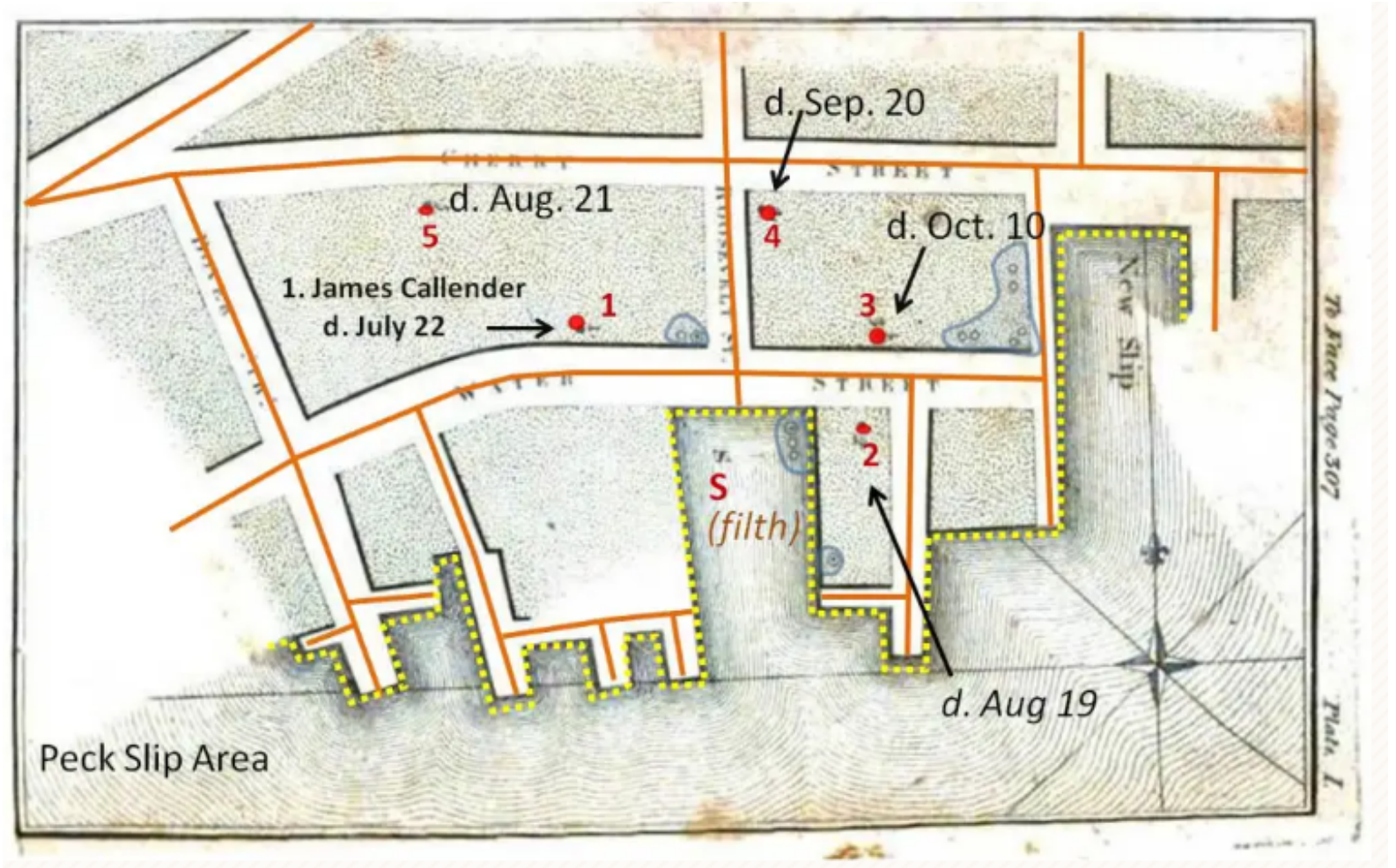
Valentine Seamen

- Made a map 1794/95!!!
- Used idea of meteorological maps



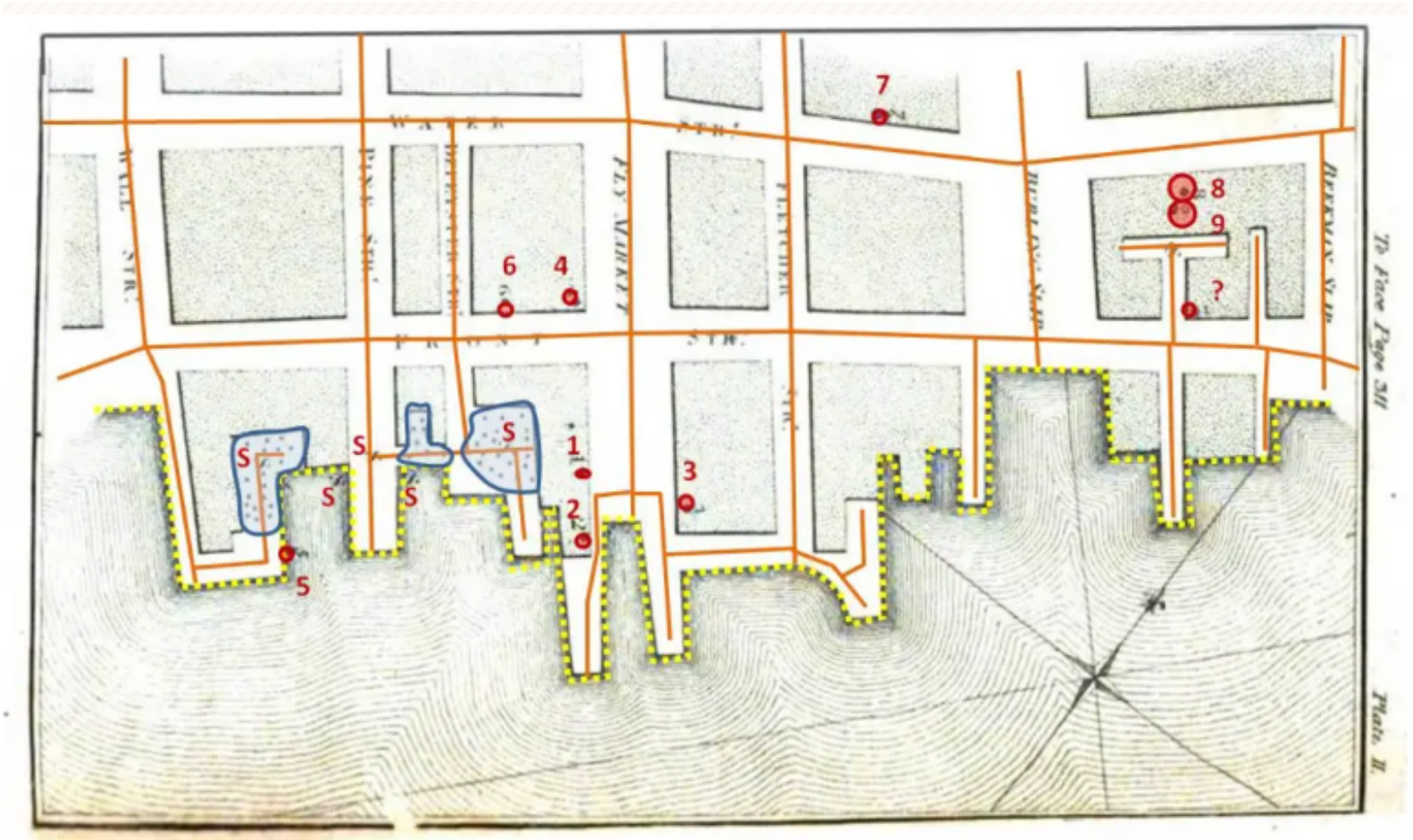
Second original map by Seamen. Source:Brian Altonen

What this map means...1



Fatal cases are noted in red, non fatal cases in slate grey. “S” stands for site of contagion or effluvium (or other miasma source). Source:Brian Altonen

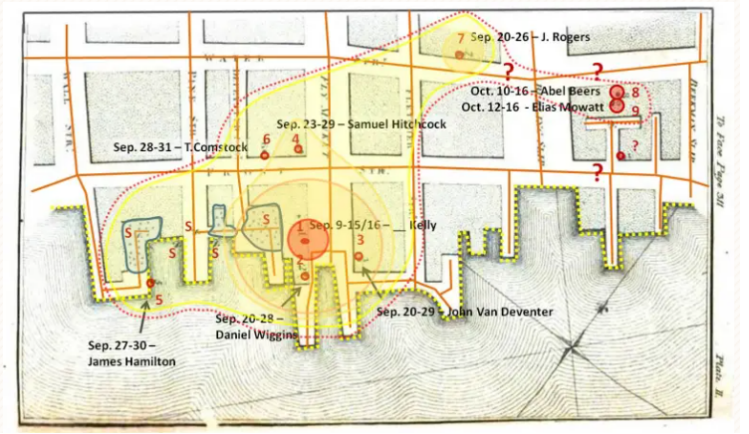
What this map means...2



Fatal cases are noted in red, non fatal cases in slate grey. “S” stands for site of contagion or effluvium (or other miasma source). Source:Brian Altonen

Interpretation

- We can see why the belief was that it was coming from ships
- There were marked **waves**
- 9th September then no cases until 20th September
- Believed to be due to
  - the time needed for decay from filth
  - ships docking
  - tidal patterns
  - slips fill and empty - when empty it would exposure the disease



Four Temporal Regions (Isopleths) with Isoline Boundaries, defined by clustering and "natural breaks". Source:Brian Altonen

Everyone was wrong as it was the Mosquito

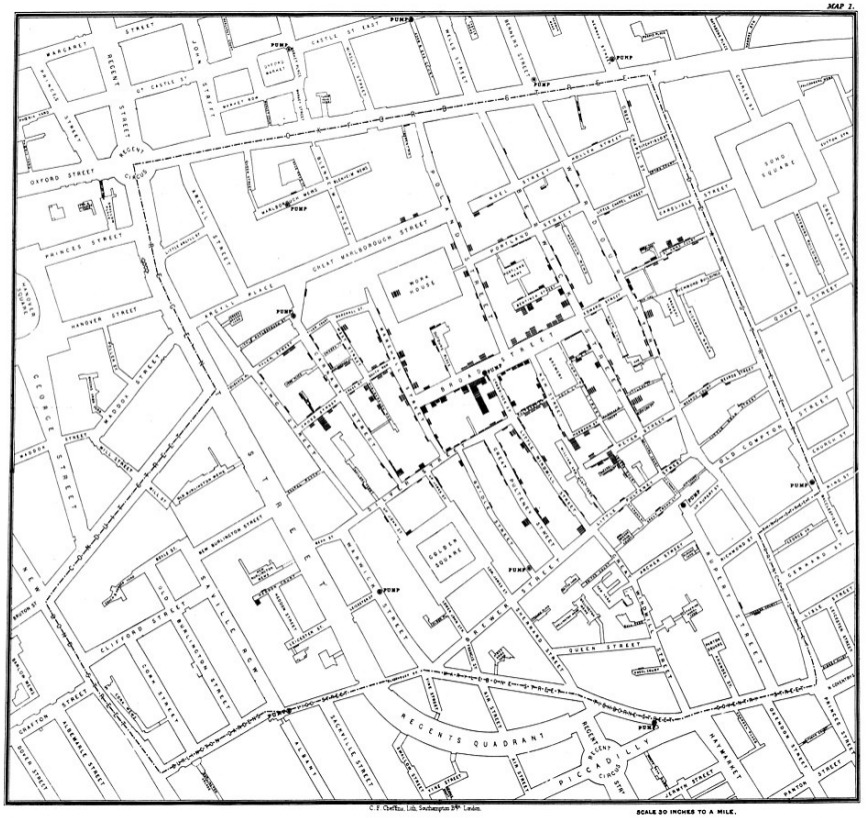
This was discounted as they thought it could be flies / other pests

Temporal nature due to breeding habbits that are related to rainfall

Correlation is not causation...

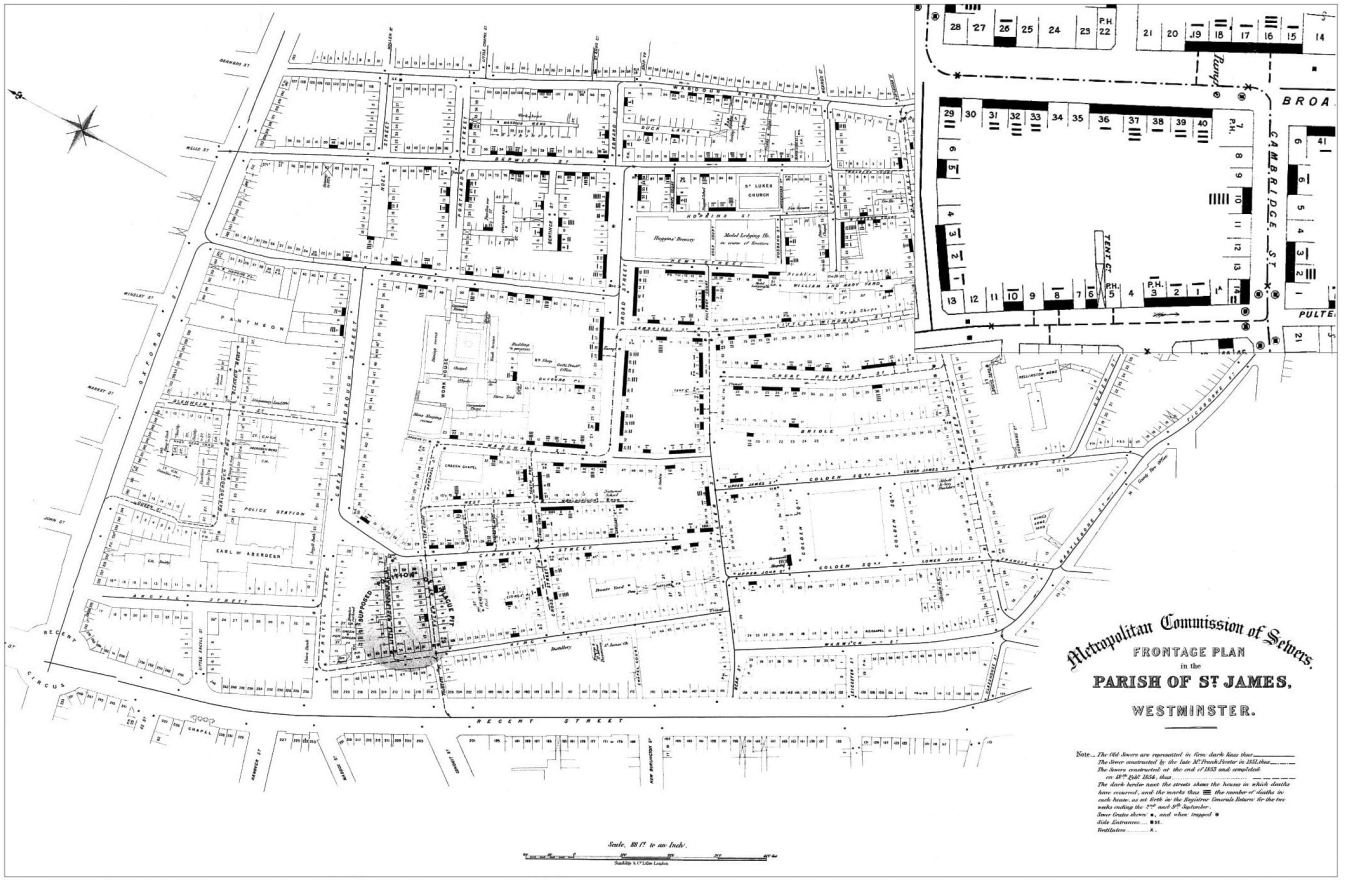
Snow's Map (?!)

- The difference here is how Snow represented the deaths - with lines
- But did Cheffins make the map ?



Something in the water: the mythology of Snow's map of cholera. Source:Kenneth Field

Map before Snow's



Deaths from cholera in Soho, London 1854 by Edmund Cooper. Source:Kenneth Field

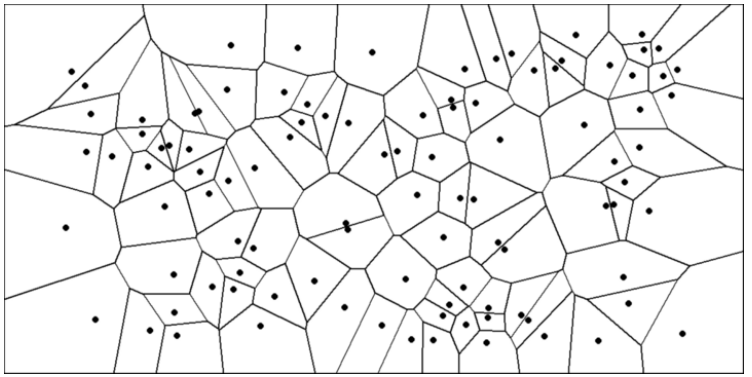
Snow's other contribution...

- Snow made two maps
- Most common one is 1854
- BUT the one in 1855 has the broad street pump in the right location
- AND also induces an isochrone - houses accessing the pump
- Is this the first kind of location analysis (in GIS location-allocation or accessibility)
- It might be the first instance of a **Voronoi Diagram**

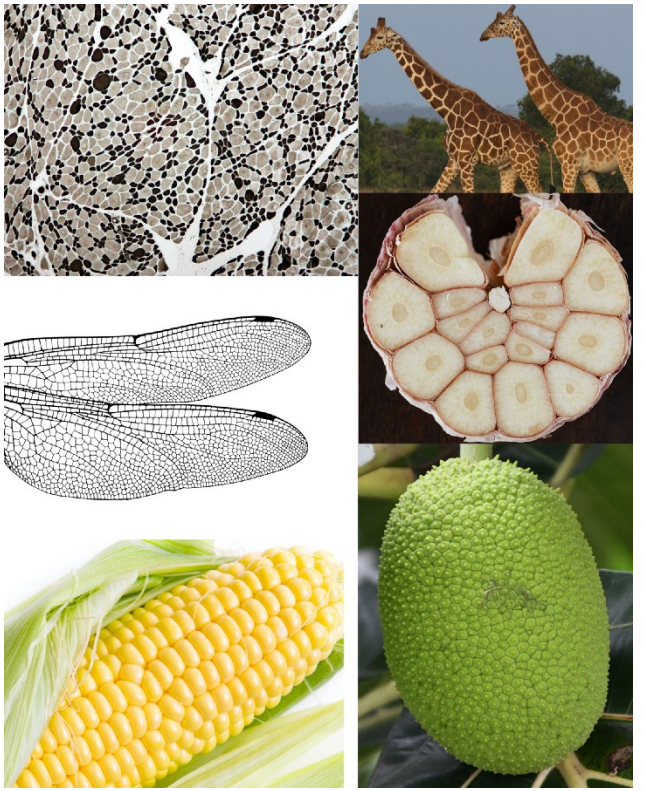
Something in the water: the mythology of Snow's map of cholera.
Source:Kenneth Field

Voronoi Diagram

- Voronoi (or Dirichlet tesselation or Thiessen polygons)
- Polygon showing the boundary of the closest point in space...



Voronoi diagram. Source:Francesco Bellelli



Voronoi in nature Source:Francesco Bellelli

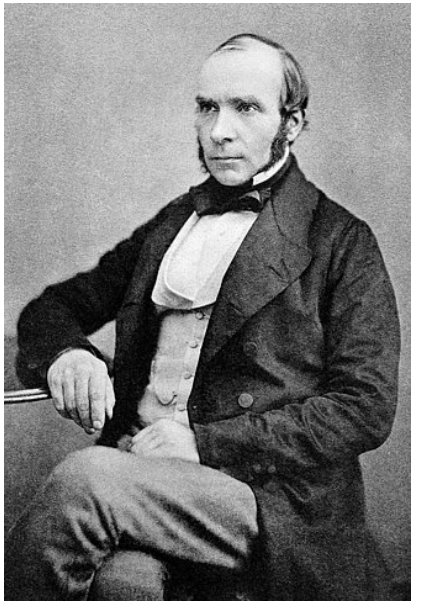
Who is responsbile?

- Valentine Seamen = 1794/95
- Edmund Cooper = 1854
- John Snow / team / cartographers = 1854

Team Seamen

Valentine Seamen ca 1800. Source:Brian Altonen

Team Snow



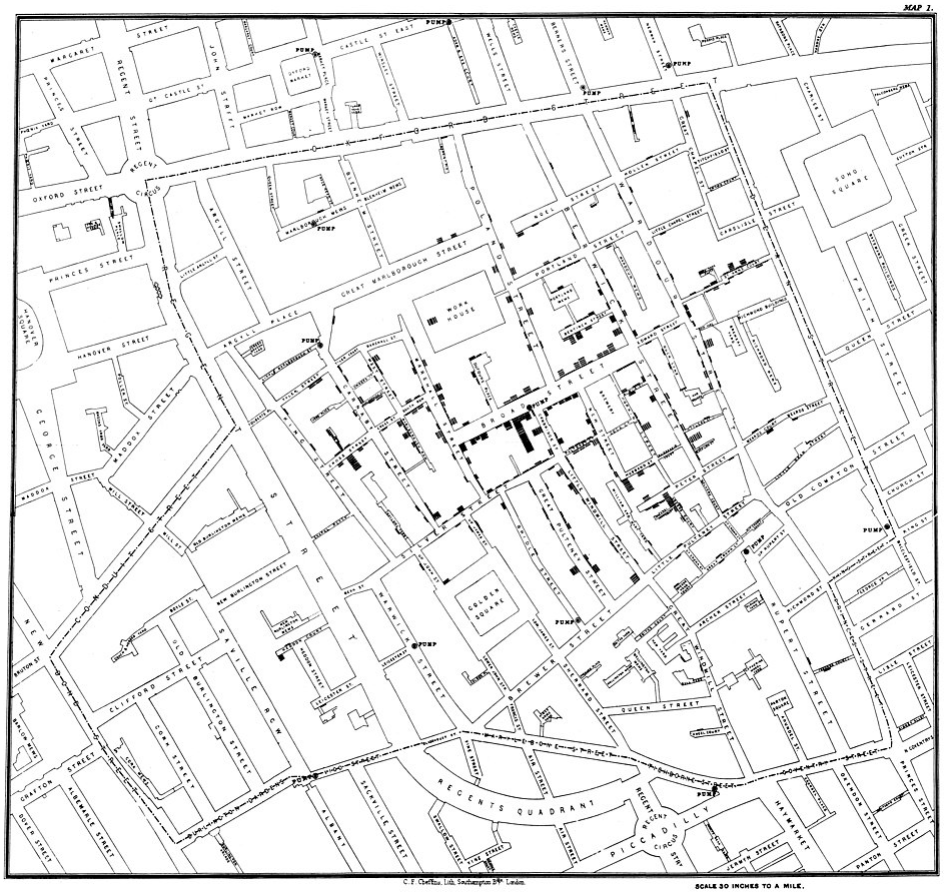
John Snow 1813-1858. Source:Wikipedia

The outcome to Snow's work

- He convinced local authority to remove the handle of the broad street pump
- By that stage it was too late
- They found a **leaking sewer** that was going into the well!!
- **Go and visit the broad street pump in London**

1854 Broad Street cholera outbreak. Source:Wikipedia

What patterns show



Something in the water: the mythology of Snow's map of cholera. Source:Kenneth Field

Spatial Epidemiology: Lung Cancer

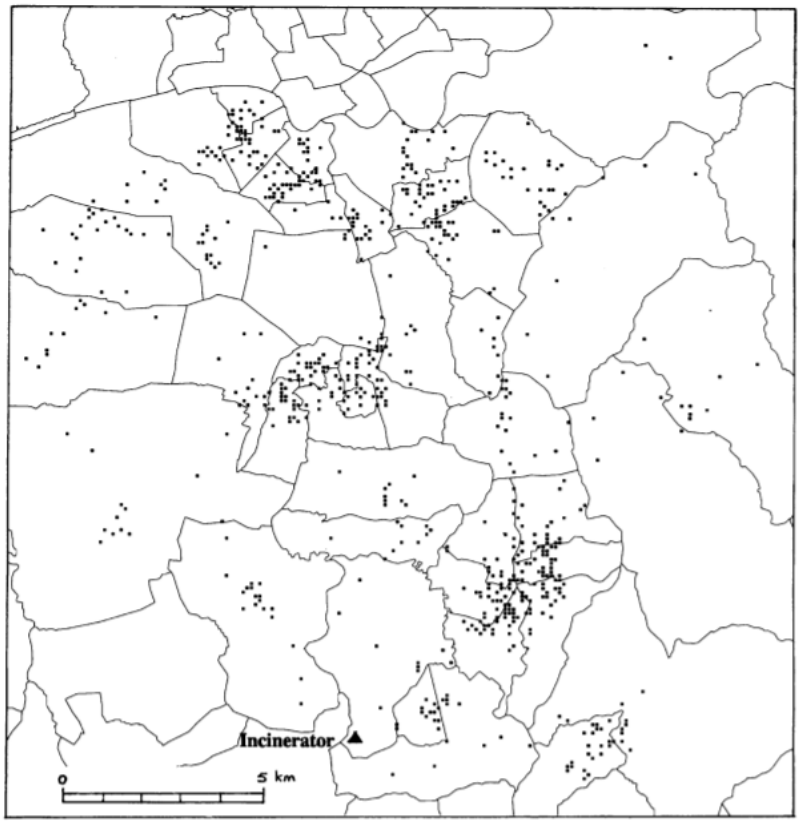


Figure 7 Locations of lung cancers, Chorley and South Ribble, Lancashire, 1974–83

Spatial Point Pattern Analysis and its Application in Geographical Epidemiology. Source:Gatrell et al.1996

Quantifying Spatial Patterns

What is fixed?

Point Pattern Analysis

- Properties are fixed (e.g. binary - present or not)
- Discrete objects - present or not, binary, yes or no.
- Examples: fly tipping, stop and search, blue plaques, pharmacies

Properties fixed, but **space (location - x,y)** can vary

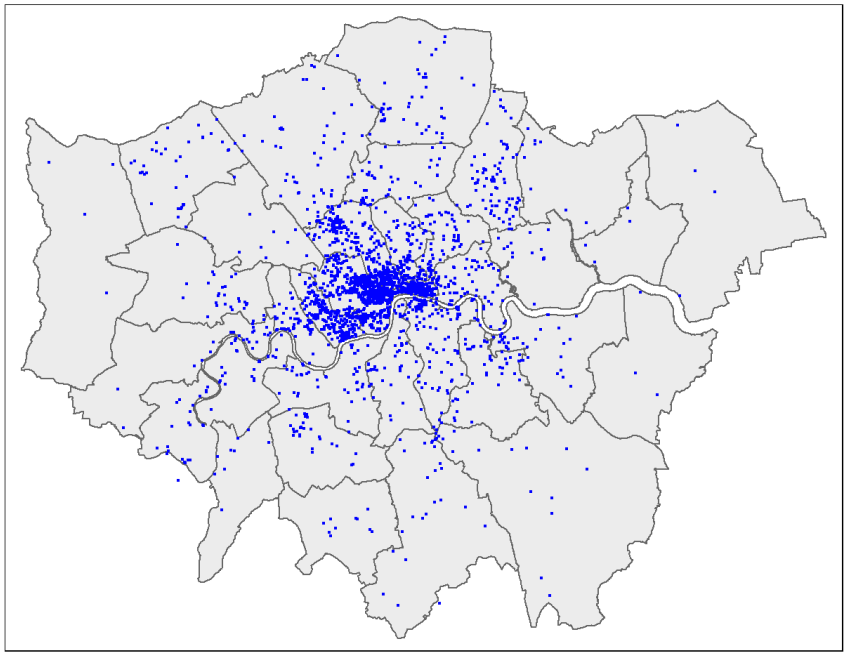
Spatial Autocorrelation

- Space (e.g. the location of the spatial units - wards, boroughs etc) is fixed
- The values of the spatial units vary
- Where the values are similar we say they exhibit Spatial Autocorrelation

Space is fixed, but **properties (values)** can vary

Examples

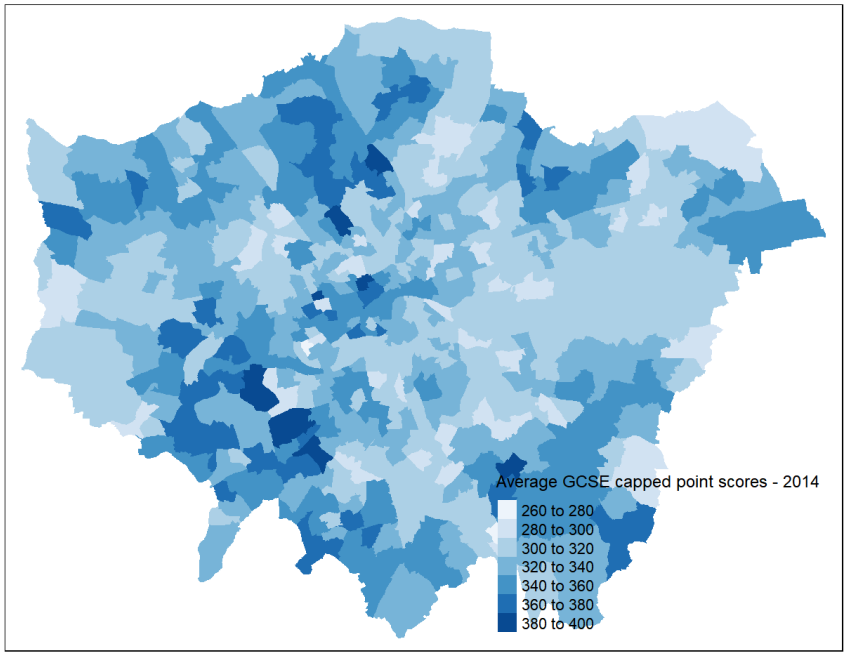
- Location of blue plaques in London
- Question: Are the points clustered or are they random?



Source:CASA0005

Examples

- Average GCSE point score 2014 per London ward
- Question: Are the values similar between certain wards?



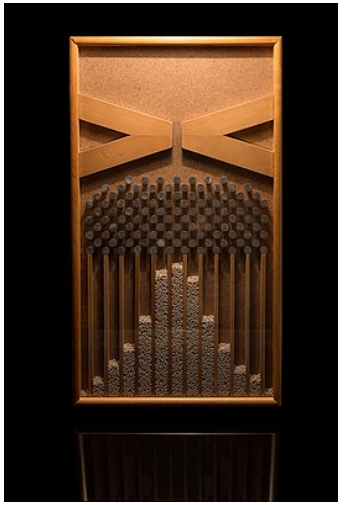
Source:CASA0005

Now how do we calculate clustering and spatial autocorrelation...

Observed vs Expected

- Comparing what we observe in the real world against what we might expect is fundamental to most spatial (and other sorts of) analyses.
- If what we observe differs in some significant way from what we might expect, then there might be something interesting going on
- But how do we know what is expected?
    - We should expect randomness
    - Randomness conforms to known probability distributions

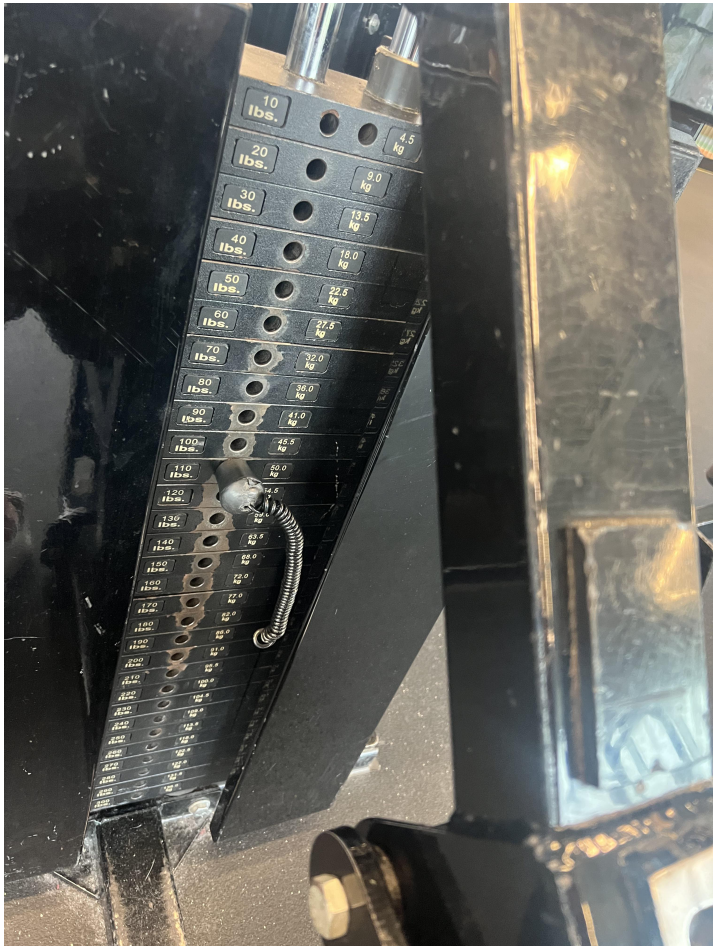
- quincunx or bean machine (or Galton box) = normal distribution



Source:Wikipedia

Francis Galton coined the term eugenics and endowed UCL with his personal collection and archive along with a bequest for the country's first professorial Chair of Eugenics. Karl Pearson was the first holder of this chair. UCL recently demaned all spaces associated with Pearson and Galton

Observed vs Expected



Source:reddit

look familiar? - the wall BBC show

Source:BBC

Eugencis at UCL

- Eugenics "the scientifically erroneous and immoral theory of “racial improvement” and “planned breeding”
- Until recently UCL had named buildings and lecture halls after these people...

Karl Perason

- Developed at UCL
- University's first Chair of Eugenics that was established on the request of Galton
- Pearson’s product-moment coefficient BUT just use **product-moment coefficient**
    - Range from -1 to 1 to show relationship between two variables....

Francis Galton

- Coined the word eugenics in 1883
- Wanted it to become a religion
- Key figure in stats but most of these were to further eugenic ideas...

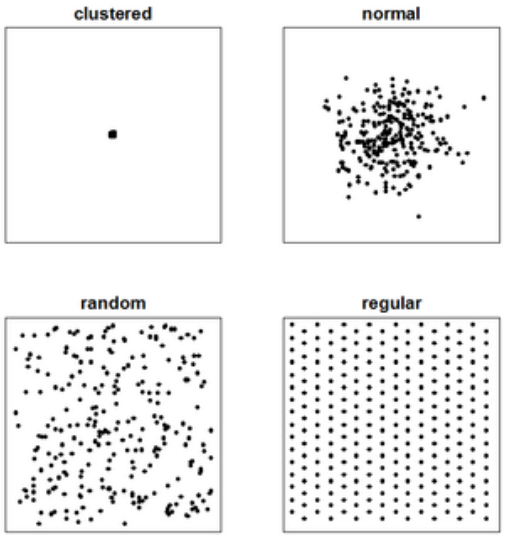
Source:Natalie Ball

Point Pattern Anaysis

The core question..

> Are these points distributed in a random way or is there some sort of pattern (uniform or clustered)?

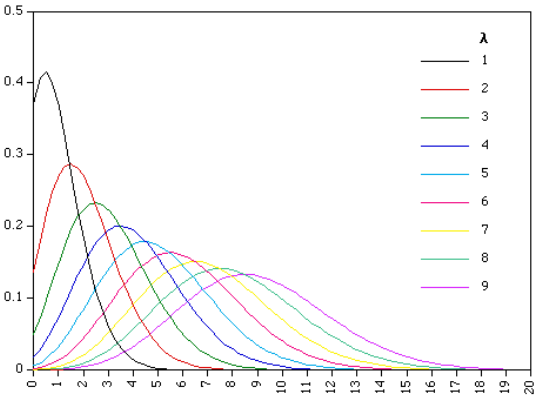
- The **expected random model** is known as Complete Spatial Randomness (CSR)
- A random distribution of points is said to have a Poisson distribution
- By comparing the distribution of observed points with a CSR Poisson model, we can tell if we have an interesting point distribution....



Source:Wikipedia

The Poisson Distribution

- Describes the probability or rate of an event happening over a fixed interval of time or space
- Where the total number of events in a fixed unit is small (e.g. Breweries in a London Borough), then the probability of getting a low rate is higher
- As number of events increases, the mean (λ – lambda) increases and the probability distribution changes



Source:UMAS

Source:Adam Dennett

The Poisson Distribution 2

The Poisson Distribution 3

Rules / applies when

- The events are discrete and can be counted in integers
- Events are independent of each other
- The average number of events over space (or time) is known

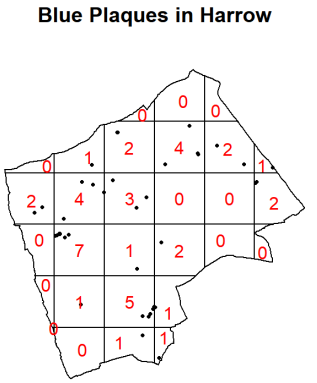
Use

- It's very useful in Point Pattern Analysis as it allows us to compare a random expected model to our observations
- Where our data **do not fit the Poisson model, then something interesting might be going on!**
- Our events might not be independent of each other – they might be clustered or dispersed and something might be causing this...

Testing for CSR - Point Pattern Analysis

Quadrat Analysis

- Developed and used frequently by ecologists
- Grid of squares
- Count number of incidents (burglaries, cholera deaths, hippos etc.) in each cell – store results in a table



Source:CASA0005

Compare the observed occurrences with a CSR Poisson model...

Note, be careful with CSR and CRS

How do we apply this to spatial data ?

In code...

For point pattern analysis we need a point pattern (ppp) object...and an observation window...

```
window <- as.owin(Harrow)

BluePlaquesSub.ppp <- ppp(x=BluePlaquesSub@coords[,1],
                          y=BluePlaquesSub@coords[,2],
                          window=window)
```

The we can create a grid...

```
BluePlaquesSub.ppp %>%
  quadratcount(.,nx = 6, ny = 6)
```

Then pull out the results...

```
Qcount <- BluePlaquesSub.ppp %>%
  quadratcount(.,nx = 6, ny = 6) %>%
  as.data.frame() %>%
  dplyr::count(Var1=Freq)%>%
  dplyr::rename(Freqquadratcount=n)
```

Quadrat Analysis

- The (Poisson) probability (Pr) of an event (brewery in a quadrat square) is calculated ... ex

$$Pr = (X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where:

- x is the number of occurrences
- λ is the mean number of occurrences
- e is a constant- 2.718
- $k!$ is the factorial of the number of occurences $((e.g. 4! = 1 * 2 * 3 * 4))$
- Here, remember $X = k$ - they are the same here.

Plugging in the numbers

Var1	Freq count	Total plaques	Lambda	Probability	Expected count	Observed probability
0	12	0	1.38	0.3	7.3	0.4
1	7	7		0.3	10.1	0.2
4	2	8		0.0	1.1	0.1
7	1	7		0.0	0.0	0.0
	29	40		1.0	28.9	1.0

- Var 1 = Number of values within the grids
- Freqquad = Number of grids with that value
- Total blue plaques = Var1*Frequency
- Lambda = Total blue plaques / total frequency
- Probability = Probability of number of plaques in quadrant ....

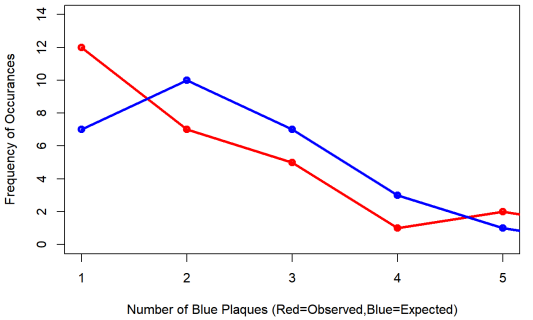
$$\frac{\lambda^k e^{-\lambda}}{k!}$$

- In excel = `(($D$2^A2)*EXP(-$D$2)/FACT(A2))`
- Expected = Expected frequency count on the Poisson distribution (freq count * probability)
- Observed probability = Frequency count / sum of the frequency count

Full example on GitHub repo in the excel document

Confirming spatial randomness

- We would expect the probability distribution of X (blue plaques in London) to have a Poisson distribution if they exhibit Complete Spatial Randomness...we can look at our plot...



- We can test for CSR by comparing the observed and expected counts and using a test such as the chi-squared
- If our p-value is > 0.05 then this indicates that we have CSR and there is no pattern in our points.
- If it is < 0.05, this indicates that we do have clustering in our points.
- Here our p-value = 0.2594, implying complete spatial randomness

What are the issues with quadrat analysis ?

(small pause)

Quadrat size

Shape

Both MAUP

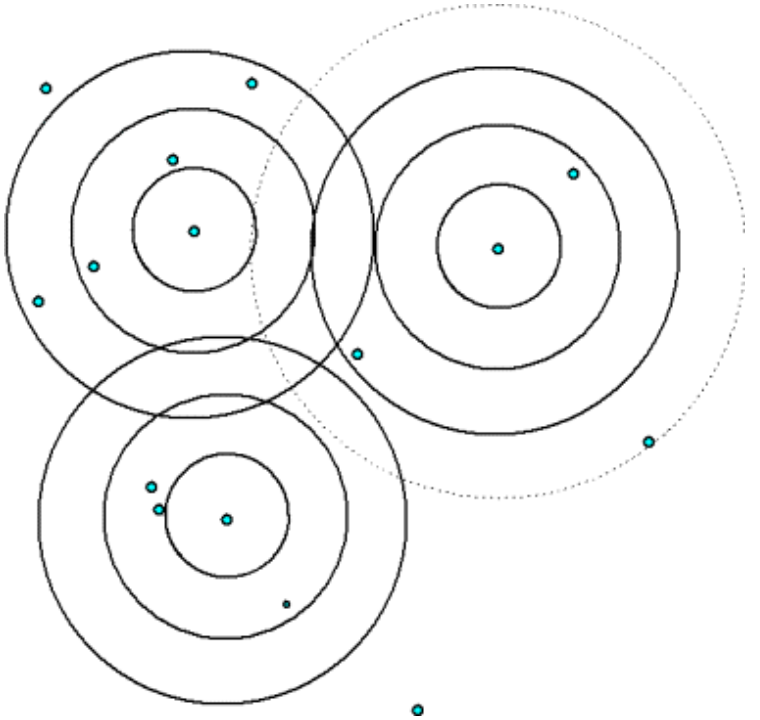
the question isn't that helpful...

Ripley's K

- To avoid scale and zoning problems associated with quadrat analysis, Ripley's K tests for CSR for circles of varying radii around each point

$$K(r) = \lambda^{-1} \sum i \sum j \frac{I(d_i j < r)}{n}$$

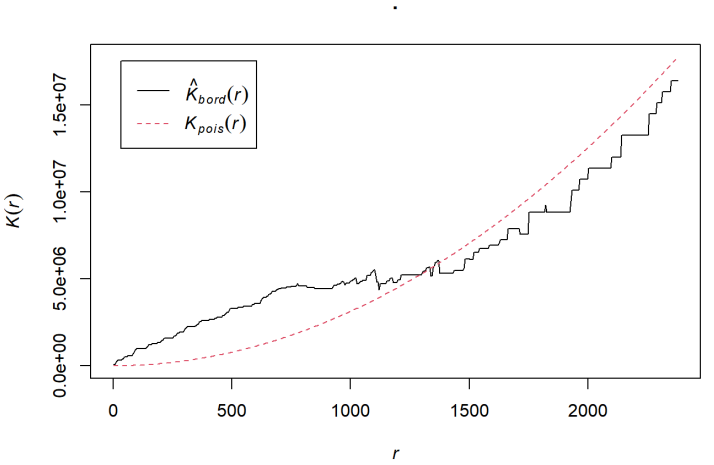
- In English: Ripley's K value for any circle radius r = the average density of points for the entire study region (of all locations) $\lambda = (n/\Pi r^2))$
- Multiplied by the sum of the distances $d_i j$ between all points within a search radius
- Divided by the total number of points, n
- I = 1 or 0 depending if $d_i j < r$



Ripley's K is simply...

```
K <- BluePlaquesSub.ppp %>%
  Kest(., correction="border") %>%
  plot()
```

- The correction specifies how points towards the edge are dealt with
- Border means that points towards the edge are ignored for the calculation but are included for the central points
- In R type `?kest` into the console for more info
- Red line is Poisson distribution (complete spatial randomness)



- Where the value of K falls above the line, the data appear to be clustered at that distance (up to 1300 metres).
- Where the value of K is below the line, the data are dispersed

- But what about planning constraints? Parks? Streets? Relative or absolute clustering?

What are the issues with Ripley's K ?

(small pause)

Computationally intensive when there are lots of points to consider

Extent of the study area can affect the calculation... eventually the radius around any point will include all other points in the study area

Phenomenon being studied cannot just occur anywhere (e.g. a river) - it's not network distance but Euclidean

We need clustering relative to other points (e.g. hotels, schools, dentists) in our study area?

We still don't know where the clusters are ?

Density-based spatial clustering of applications with noise (DBSCAN)

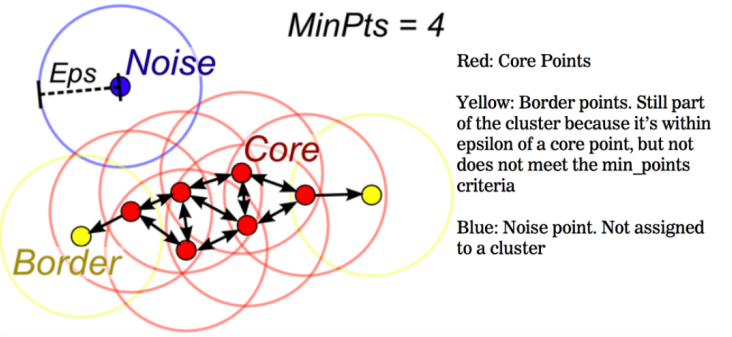
Popular because can detect non-linear clusters

2 parameters:

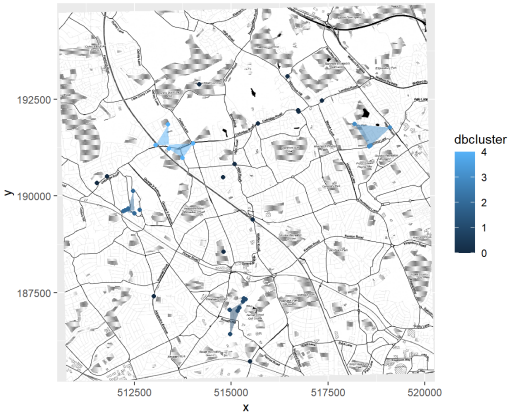
- Epsilon (eps) = size of neighbourhood within which to search for other points
- Minimum number of points to search for (MinPts)

If a point has >= MinPts in neighbourhood then defined as 'Core'

If point in neighbourhood of a core point but has <MinPts in its own neighbourhood, then defined as 'Border'



DBSCAN. Source:Medium



Source:CASA0005

DBSCAN

In R DBCAN is simply...

```
db <- BluePlaquesSubPoints %>%
  fpc::dbscan(.,eps = 700, MinPts = 4)
```

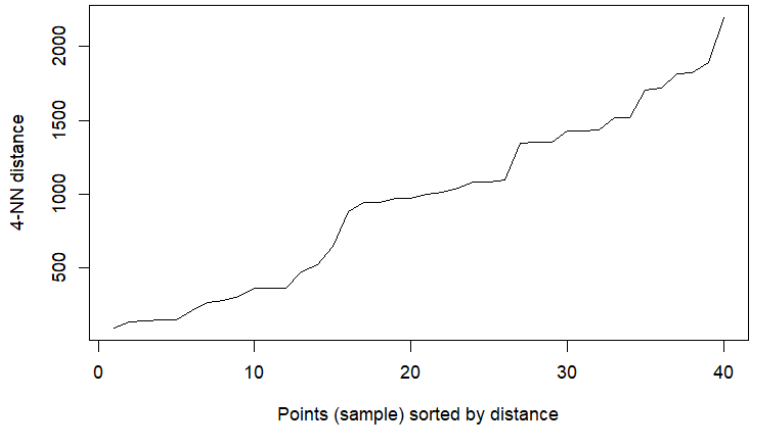
How to select values

- Eps: use the Ripley's K buldge - where the line shows the greatest difference to the theoretical value from a few slides ago...around 700 meters
- Or plot the **average distance of each point to k neighbours**, which are then plotted in ascending (low to high) order.
    - The knee is where this value (of distance to neighbours) increases and the value we use.

```
BluePlaquesSubPoints%>%
  dbscan::kNNdistplot(.,k=4)
```

DBSCAN

...this gives...



- plotted: average distance to the k neighbours, which are then plotted in ascending order
  - The knee is where this value (of distance to neighbours) increases and the value we use
- Min points: start with what you think is appropriate
- OPTICS - extends DBSCAN and uses an algorithm to find optimum distance thresholds for clustering.

HDBSCAN - not in practical

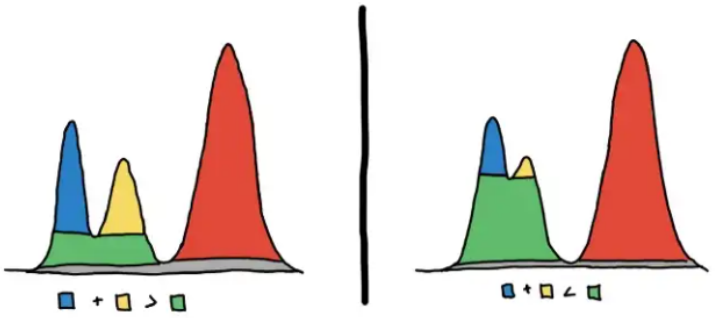
- Hierarchical Density-based Spatial Clustering of Applications with Noise

Other methods good when:

- round data, equally sized, equally dense, no noise.

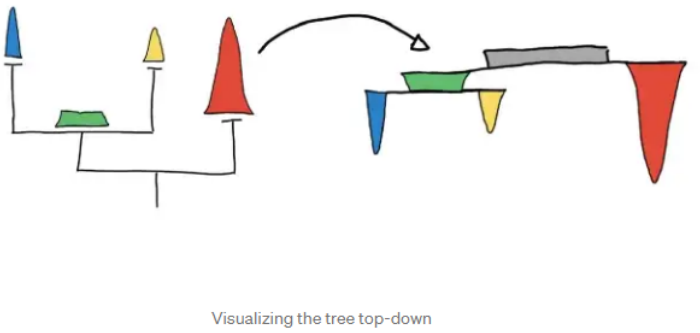
HDBSCAN

- cluster hierarchy, looks for regions of dense data (not shape / eps / distance can vary)
- e.g. islands from the sea ...



3 clusters on the left vs 2 clusters on the right

HDBSCAN. Source:Pepe Berba

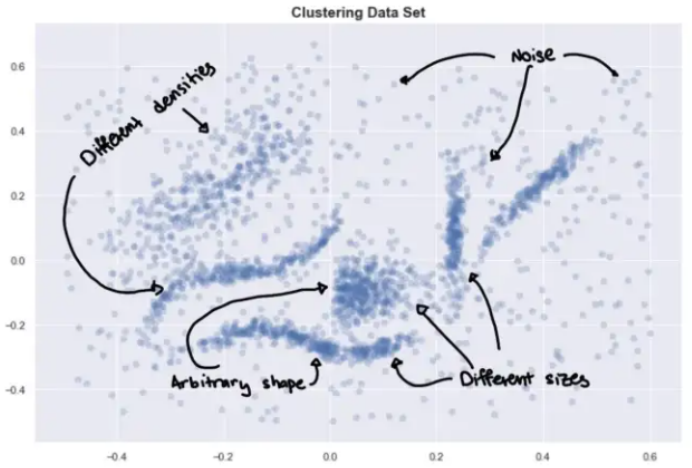


Visualizing the tree top-down

HDBSCAN. Source:Pepe Berba

HDBSCAN...best...

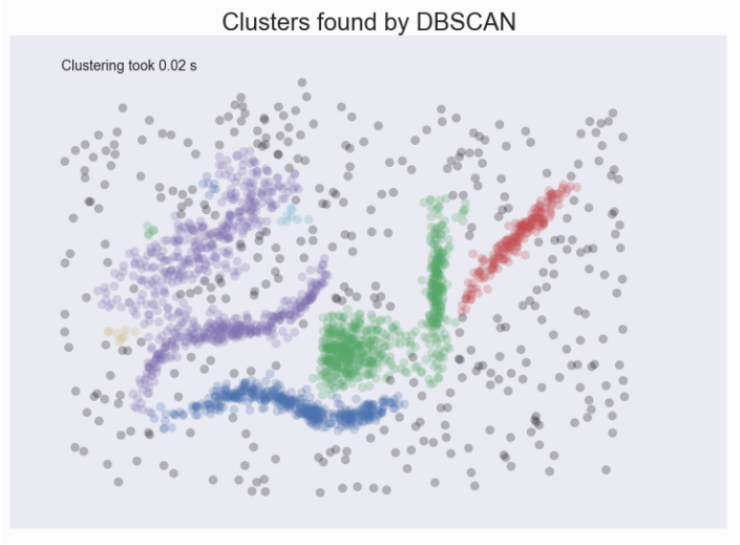
- Arbitrary shapes - not all circular
- Clusters of different sizes - not constrained by 1 argument (e.g. distance or K neighbours)
- Clusters with different densities
- Noise / outliers



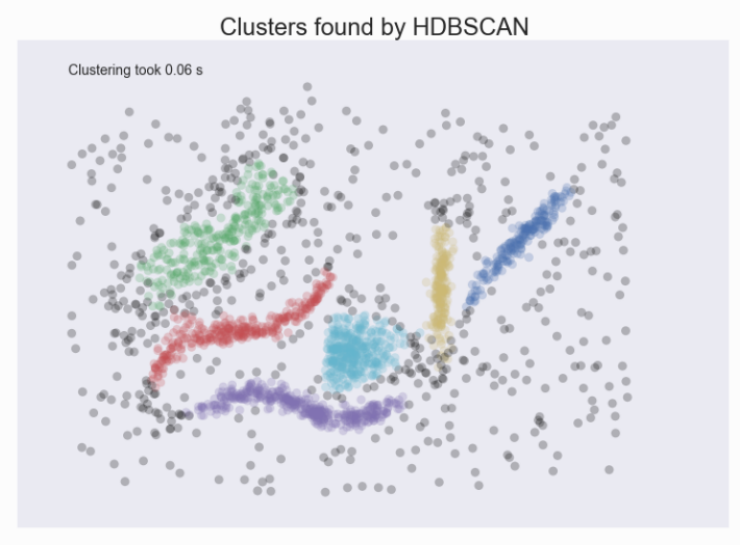
HDBSCAN. Source:Pepe Berba

HDBSCAN...differences...

- From the same authors that created DBSCAN
- it is the same but...**there is no epsilon (distance requirement)** and it can **VARY**



Source:Comparing Python Clustering Algorithms



Source:Comparing Python Clustering Algorithms

The core question..

> Are these points distributed in a random way or is there some sort of pattern (uniform or clustered)?

Adam's paper on beer...

GJ The Geographical Journal

The Geographical Journal, Vol. 183, No. 4, December 2017, pp. 440–454, doi: 10.1111/geoj.12228

The geography of London's recent beer brewing revolution

ADAM DENNETT* AND SAM PAGE†
*Centre for Advanced Spatial Analysis, University College London, Gower Street, London WC1E 6BT
E-mail: a.dennett@ucl.ac.uk
†Department of Geography, University College London, Gower Street, London WC1E 6BT
E-mail: s.page.12@ucl.ac.uk
This paper was accepted for publication in July 2017

In this paper we examine the recent rapid growth of new breweries in London and the reasons behind it. At the turn of the millennium, just a handful of breweries was operating in London, but by 2016 this number had risen to over 85. Using open data from the Companies House database augmented with other online and printed sources, we show that the rapid growth of breweries, particularly since 2011, has exhibited spatial patterning. Ripley's K analysis reveals that as soon as we see new breweries emerging, they are clustering in space. Cluster analyses reveal that Bermondsey and Hackney are particular locational hotspots for brewing. Closer investigation of the Bermondsey cluster highlights the importance of a number of interacting physical, social and economic factors in helping foster this growth. We show that the railways and the spaces they have created, the general atmosphere of cooperation and sharing surrounding the industry in the city, macro-economic and fiscal changes, foreign influence, technology and markets have all played their part in the recent spatial and temporal evolution of brewing in the city.

KEY WORDS: London, brewing, spatial analysis, beer

Introduction

In the 50 years between the mid-1920s and the mid-1970s, the number of breweries operating in the UK fell from some 2000 to just 87 (Dixon 1978; SIBA 2010). In the 1970s and 1980s further closures, takeovers and mergers, combined with changing tastes and drinking habits of the beer drinking populace[1], meant that some of the most famous and well established names in British beer, such as Truman's of London, disappeared. By the end of the 1980s, brewing in the UK was dominated by just a handful of large breweries (Knowles and Egan 2001) producing an increasingly homogenised product. The Campaign for Real Ale (CAMRA) was established in 1971 to try to arrest the decline of the once great British pint, but for many beer drinkers the closure of so many breweries meant the historic diversity and variety in British beer and brewing had been replaced by bland homogeneity.

However, anyone buying a pint of beer in a British pub in recent years may have noticed that the ubiquitous brands of the large multinational beer conglomerates have slowly started to share tap space with a growing variety of alternative producers. It is fair to say that the UK brewing industry has undergone something of a renaissance in recent times. Data from the Society of Independent Brewers (Cabras 2015) have shown that beer production by its members almost doubled from 2009 to 2014, from 1,721,291 hectolitres to 2,992,747. Data from Companies House (which records all active and some dormant and inactive companies in the UK) documents an almost exponential growth in 'Manufacturers of Beer', with a 13-fold increase in active breweries in the last 20 years.

While new breweries have sprung up all over the UK, London has played a significant role. Despite its rich brewing history (Brown 2015; Krenzke 2014), by the early 1990s only a handful of brewers remained in London[2] and of those, only two survived into the second decade of the twenty-first century – Fuller Smith and Turner (or Fuller's) located at the Griffin Brewery in Chiswick on the north bank of the Thames; and Young's, on the South Bank of the river at the Ram Brewery in Wandsworth[3]. Of the London breweries operating today only the Meantime Brewery[4] in Greenwich and Fullers were operating before the turn of the millennium.

- Is there anything interesting about the spatial distribution of Breweries in London?
- Hypothesis: If terroir [environmental factors] and local provenance confer brand [stakeholders] (and financial) advantage, then dispersal is likely. If sharing of knowledge, resources (hardware, recipes, staff, customers) more important, then clustering likely.

But... how similar is our data across geographic space?

sort of.. we are now considering spatially continous data (e.g. density of the points) as opposed to the points themselves...

But when we move to spatial units ...be careful...as Gimmond 2022 states...

> you might be tempted to state that there is a strong relationship between variables v1 and v2 at the individual level. But doing so leads to the ecological fallacy where the statistical relationship at one level of aggregation is (wrongly) assumed to hold at any other levels of aggregation (including at the individual level). In fact, all you can really say is that “at this level of aggregation, we observe a strong relationship between v1 and v2” and nothing more!

Part 2: Spatial autocorrelation

Questions we can ask / set

Points

Are these points distributed in a random way or is there some sort of pattern (uniform or clustered)?

Spatially continuous observations (e.g. values of polygons)

How (dis)similar are our values assigned to geographic units across geographic space

Remember - what is fixed ?

Quantifying Spatial Patterns

What is fixed?

Point Pattern Analysis

- Properties are fixed (e.g. binary - present or not)
- Discrete objects - present or not, binary, yes or no.
- Examples: fly tipping, stop and search, blue plaques, pharmacies

Properties fixed, but **space (location - x,y)** can vary

Spatial Autocorrelation

- Space (e.g. the location of the spatial units - wards, boroughs etc) is fixed
- The values of the spatial units vary
- Where the values are similar we say they exhibit Spatial Autocorrelation

Space is fixed, but **properties (values)** can vary

Just because we have clustering doesn't mean there will be patterns of spatially referenced continuous observations

Example - Spatial Epidemiology: Mortality

- Similar values might suggest there is something more going on
- Some sort of spatial influence
- Note the difference between spatially continuous and point data

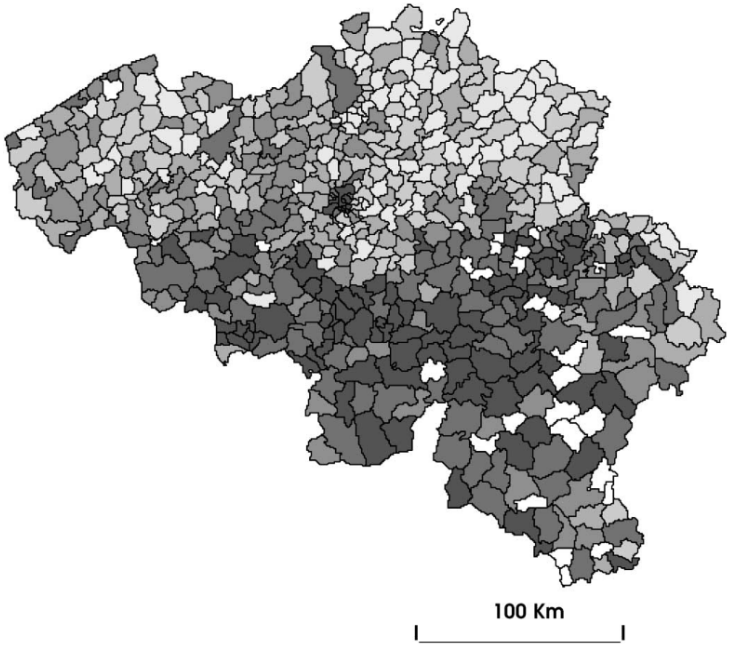
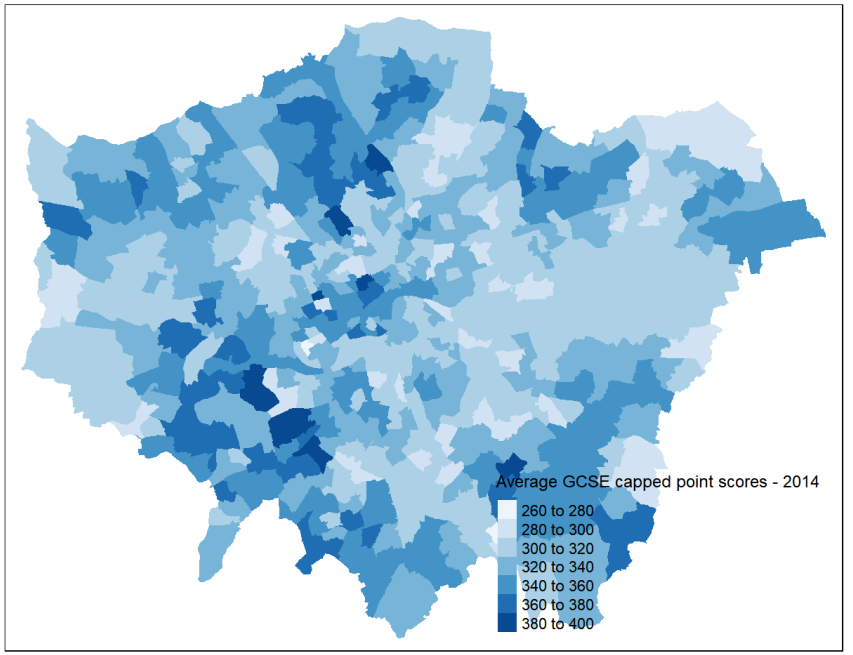


Fig. 1. Mortality ratio increases from light grey (low mortality) to black (high mortality). Quintile method of categorisation used White municipalities have missing information.

Mortality ratio increases from light grey (low mortality) to black (high mortality). Quintile method of categorisation used White municipalities have missing information..
Source:Lornt et al.2001

Example - GCSE scores

- Average GCSE point score 2014 per London ward
- Question: Are the values similar between certain wards?



Source:CASA0005

From points to spatially continuous observations 1

- We can convert almost any point data into spatially continuous observations

Questions

- Why should we never (or almost never) use count data (e.g. number of blue plaques per ward) for spatial autocorrelation measures?
- In what instance might we be able to use count data?

How would you go from point data to continuous observations

What are spatially continuous observations ?

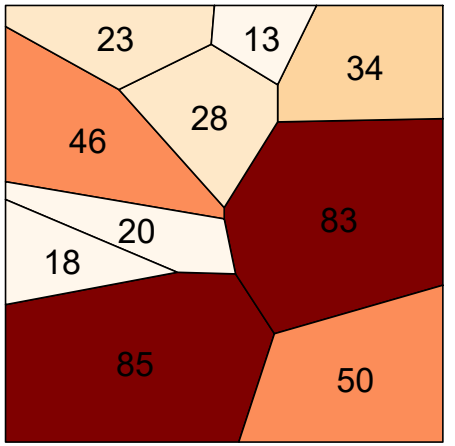
continuous observations over a surface e.g. temperature

From points to spatially continuous observations 2

Answers

- Summing the counts generates results that are dependent on the size of the spatial unit ...e.g...

16	16	16	16	16
16	16	16	16	16
16	16	16	16	16
16	16	16	16	16
16	16	16	16	16



Count of individuals in each zonal unit. Note how an underlying point distribution can generate vastly different looking choropleth maps given different aggregation schemes. Source:Intro to GIS and Spatial Analysis Manuel Gimond, 2022

From points to spatially continuous observations 3

- In this case we could sum our counts and normalise then based on some sort of underling data - such as population. e.g...lung cancer rates or mortality rates

Our code would look similar to this ...

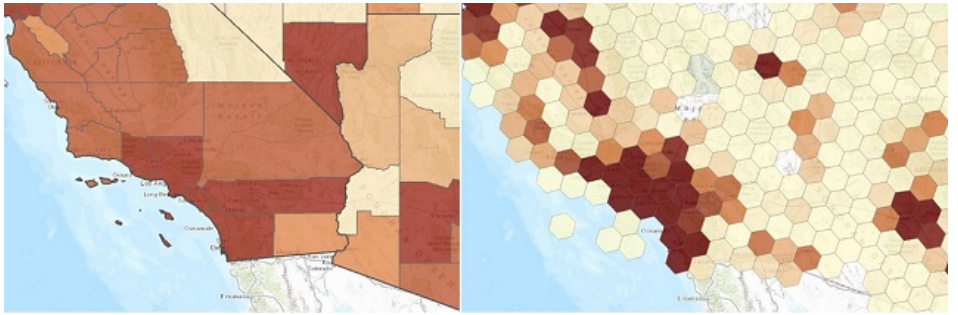
```
points_sf_joined <- LondonWardsMerged%>%
  st_join(BluePlaquesSub)%>%
  add_count(ward_name)%>%
  janitor::clean_names()%>%
  #calculate area
  mutate(area=st_area(.))%>%
  #then density of the points per ward
  mutate(density=n/area)%>%
  #select density and some other variables
  dplyr::select(density, ward_name, gss_code, n, average_gcse_capped_point_scores_2014)
```

- Sometimes we can have "unstable rates" - where values are very small, consult Coping with Unstable Rates

From points to spatially continuous observations 4

Answers

- An instance where it is ok to use counts is where the mapping units are spatially consistent --e.g. hexagons or Uber's H3 grid
- The variable spatial unit is removed

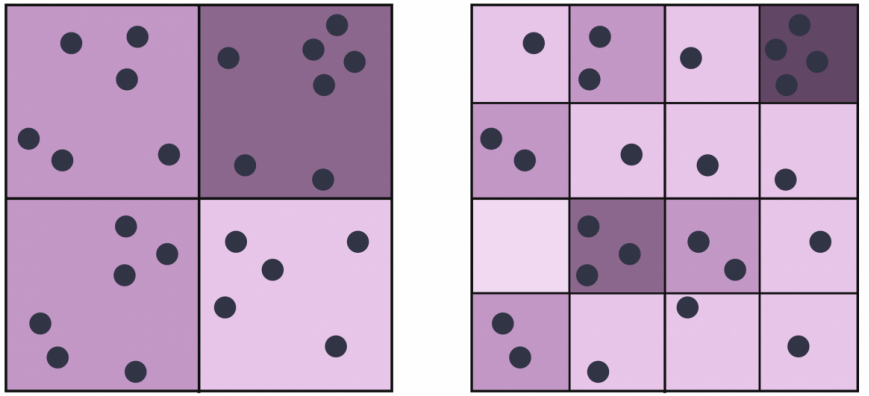


Source:Thematic Mapping with Hexagons, ESRI 2015

Take care!

Modifiable Areal Unit Problem

- Different aggregations will lead to:
  - Different outcomes
  - Incomparable results



Source:gisgeography

Take care 2!

If you have aggregated point data they can no longer be described as points...as Gimond 2022 says

> "you might be tempted to state that there is a strong relationship between variables v1 and v2 at the individual level. But doing so leads to the ecological fallacy where the statistical relationship at one level of aggregation is (wrongly) assumed to hold at any other levels of aggregation (including at the individual level). In fact, all you can really say is that “at this level of aggregation, we observe a strong relationship between v1 and v2” and nothing more!"

Spatially continuous observations

Tobler's First Law of Geography

fundamental to spatial analysis underpins

- spatial dependence (values vary over space - e.g. weather station point data, depends on the location and on other values)
- autocorrelation (in this lecture)
- interpolation (estimating values on a surface e.g. weather station point data)

> "Everything is related to everything else, but near things are more related than distant things."

Waldo R. Tobler. Source:Wikipedia

Tobler W., (1970) "A computer movie simulating urban growth in the Detroit region". Economic Geography, 46(2): 234-240.

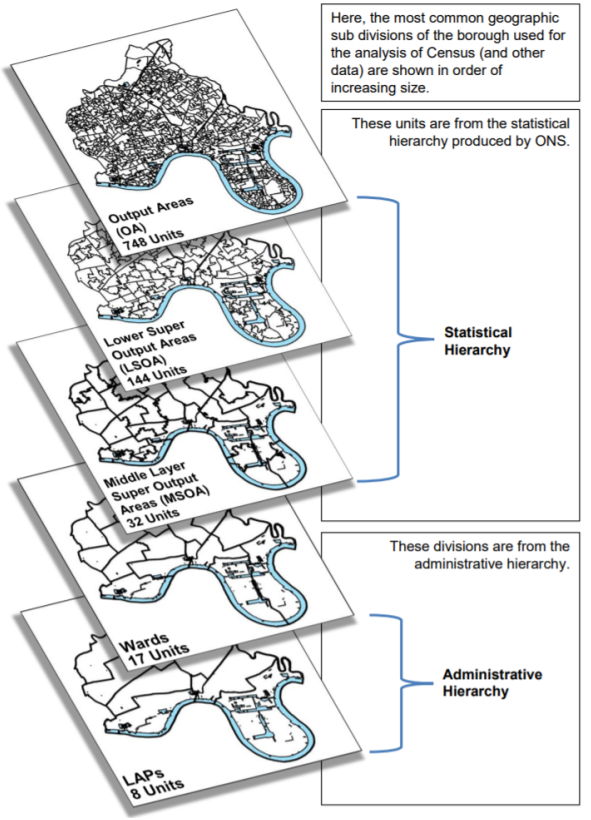
But ...

"Everything is related to everything else, but near things are more related than distant things" always holds absolutely.

> This is and has always been an oversimplification, disguising possible underlying entitation, support and other mis-specification problems. Are the units of observation appropriate for the scale of the underlying spatial process? Could the spatial patterning of the variable of interest for the chosen entitation be accounted for by another variable?, Edzer Pebesma, Roger Bivand

Spatially continuous observations

But this does depend on our level of aggregation here...



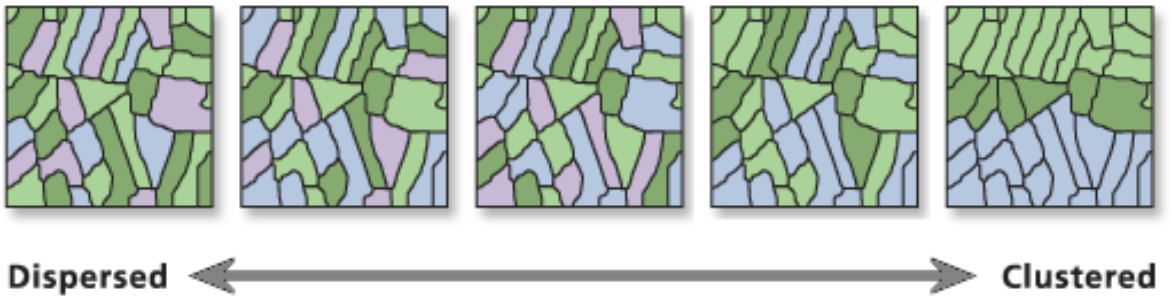
Spatial Units. Source: Tower Hamlets

Spatially continuous observations

Formal definition:

The (auto*)correlation among observations of a single variable (smoking rates, unemployment, etc.) that is strictly attributable to the proximity of those observations in geographic space.

*Auto = with itself

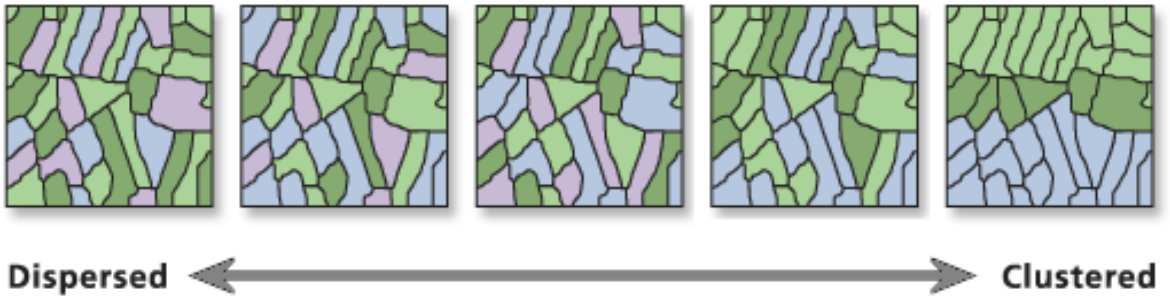


Moran's I. Source:ArcGIS

Spatially continuous observations

Interpretation:

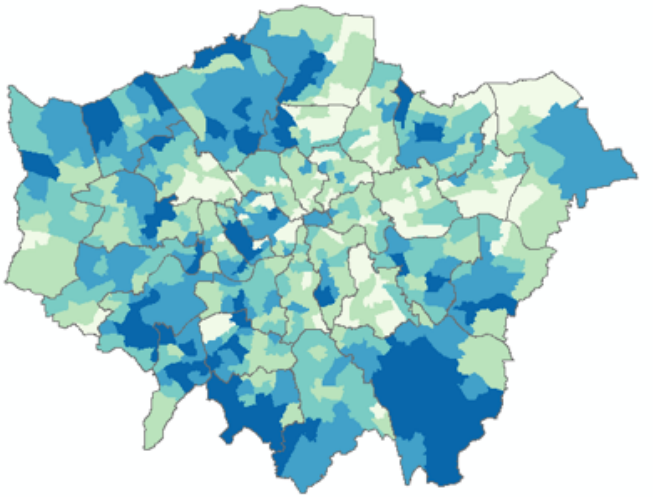
- Are the observations (data)
- Similar to those (of the same data) near (proximity) them or different over geographic space



Moran's I. Source:ArcGIS

Spatial Autocorrelation

- Frequently in spatial analysis we don't just want to study discrete events, **but the ways in which variables change across space...**
- Everything is related to everything else, but near things are more related than distant things
- Are the GCSE scores of pupils in London more likely to be **similar in areas that are close to each other than those in areas that are distant**?
- Similar observations in similar places might be the result of some underlying cause



- Are the values clustered or are they random?

What algorthims do we use to check for clustering of spatially continuous data?

The can be global (single value)

Or local (values change over space) = Local Indicators of Spatial Association (LISA)

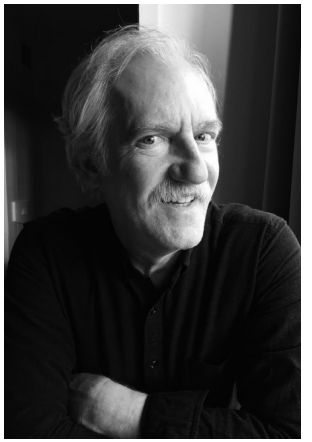
The players...

Left to right

- Pat Moran (1917-1988) = **Moran's I** and then Local Moran's I by Luc Anselin
- Arthur (Art) Getis (1934-2022) = **Getis Ord General G**
- Keith Ord (? - still works i think at Georgetown uni) = **Getis Ord General G**
- Robert (Roy) Charles Geary (1896-1983) = **Geary's C**

Image below

- Luc Anselin (1953-present) = **Anselin Moran's I (local)**



Indices of Spatial Association

- Moran's I, Geary's C and Getis-Ord's G are all indexes to compare values for neighbouring features
- Help answer the next core question –

are the values for neighbouring features more similar than those for all other features

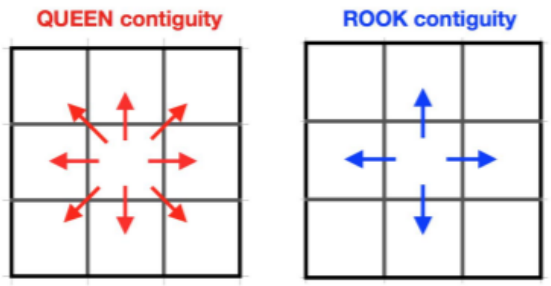
if the average difference between neighbouring features is less than between all features, values are clustered

How do we define a neighbour to a polygon ?

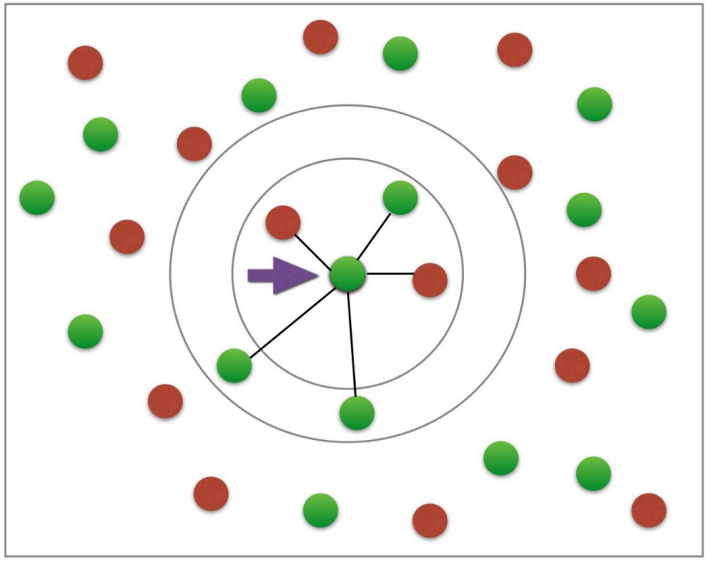
In the classroom identify who is your neighbour?


Neighbours

What case did you use?



Neighbour cases. Source:Bellefon and Gleut, INSEE



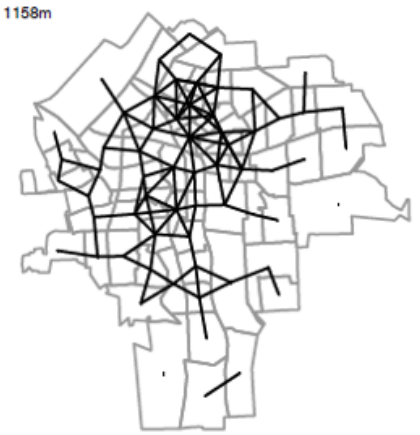
Nearest neighbour. Source:Towards Data Science

What case did you use?

- Are the values in near places similar or dissimilar
  - **Adjaceny** = what is next to my polygon should i compare to. **k** nearest (e.g. k=4, means 4 nearest from the centroid of the polygon)

What case did you use?

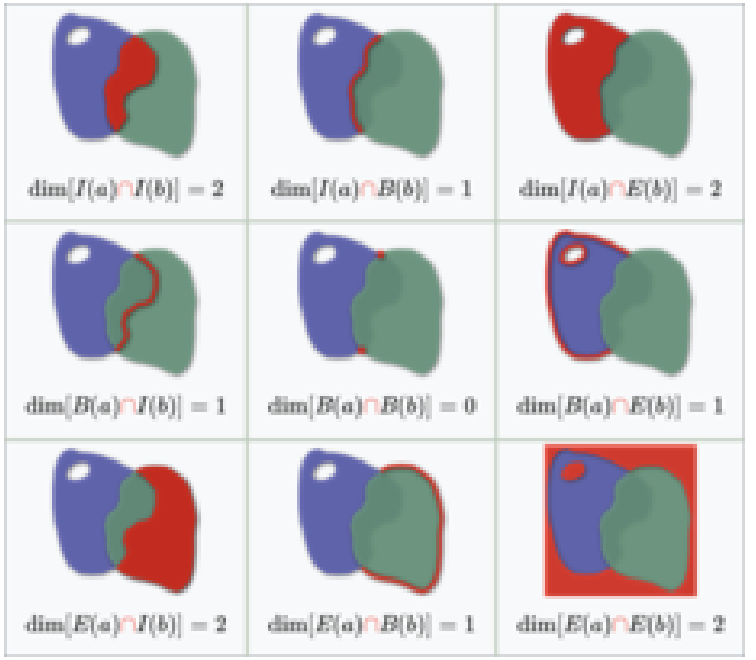
- Are the values in near places similar or dissimilar
  - **Distance** = how many polygons within a set distance should i compare to (from the centroid of the polygon)



What case did you use?

topology or topological relations = the spatial relationships between adjacent or neighboring features

- The Dimensionally Extended 9-Intersection Model (DE-9IM) is the standard topological model used in GIS to define spatial relationships between objects
- If boundaries touch in some way, then we would usually classify objects as being neighbours



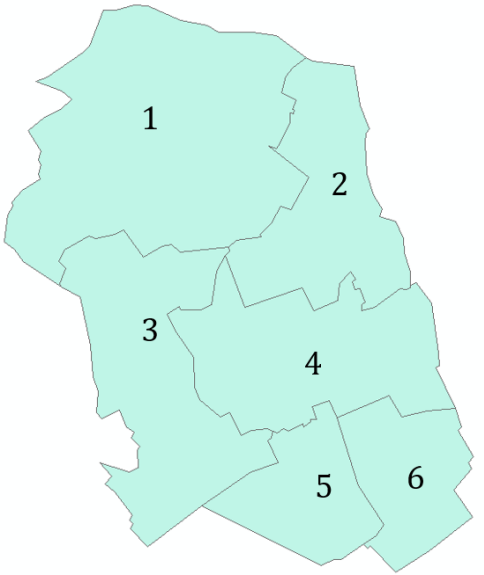
DE-9IM. Source:Wikipedia

How do we show what is a neigbour in GIS?

Spatial weight matrix

- A **binary (yes/no) matrix**

W_{ij}	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6
Zone 1		1	1	0	0	0
Zone 2	1		1	1	0	0
Zone 3	1	1		1	1	0
Zone 4	0	1	1		1	1
Zone 5	0	0	1	1		1
Zone 6	0	0	0	1	1	



Weight matrix. Source:Adam Dennett

- In matrix algebra, it is convention to refer to **rows** with the index, i and **columns** with the index, j
- W_{ij} refers to any weight in this matrix
- W_{13} =1, W_{52}=0 etc.

The complete process

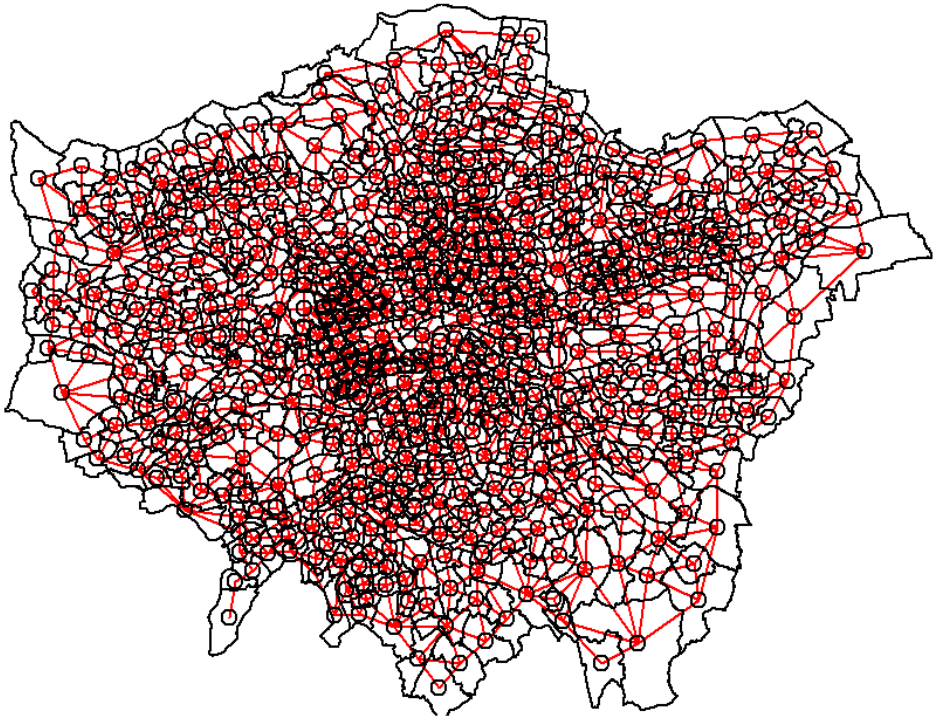
1. Set **the rule** - what is a neighbour: Rook, Queen, KNN or even just a distance for a list of neighbours

```
#create a neighbours list
LWard_nb <- points_sf_joined %>%
  poly2nb(., queen=T)
```

1. Set how **the matrix is populated** - gives a numeric value to the neighbour.
    - B is the basic binary coding (1/0) (previous slide)
    - W is row standardised (sums over all links to n)
    - C is globally standardised (sums over all links to n)

```
#create a spatial weights matrix from these weights
Lward.lw <- LWard_nb %>%
  nb2mat(., style="B")
```

GIS what is a neighbour 2



Standardisation of matrix

Standardisation **permits comparable** spatial parameters where features might be biased due to the aggregation...

When you standardise (usually row) the fact that one unit has 20 neighbours and another 2 doesn't influence the results that much, but it would in binary. When we aggregate data the structure (e.g. polygons) aren't made for (or designed for) our point data (e.g. blue plaques, fly tipping etc). See this ESRI help link

Global (matrix from previous slides)

- Sum the weights = 18
- Divide our units by this = 6/18
- Each weight is given 0.33

Row

- 1 is divided by the sum of the number of neighbours in each row
- In zone 1 each weight is 0.5 - Binary it would sum to 2
- In zone 3 each weight is 0.25 - Binary it would sum to 4

Example of Moran's I (global)...1

1. Work out the matrix....

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

First part (left)

n	Number of zones						
Bottom pa	Sum of weights						
Binary							
	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Sum
Zone 1	0	1	1	0	0	0	2
Zone 2	1	0	1	1	0	0	3
Zone 3	1	1	0	1	1	0	4
Zone 4	0	1	1	0	1	1	4
Zone 5	0	0	1	1	0	1	3
Zone 6	0	0	0	1	1	0	2
J=6						Sum of weights	18
Row standardised							
	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Sum
Zone 1	0	0.5	0.5	0	0	0	1
Zone 2	0.33333	0	0.33333	0.33333	0	0	1
Zone 3	0.25	0.25	0	0.25	0.25	0	1
Zone 4	0	0.25	0.25	0	0.25	0.25	1
Zone 5	0	0	0.33333	0.33333	0	0.333333333	1
Zone 6	0	0	0	0.5	0.5	0	1
N=6						Sum of weights	6

Example of Moran's I (global)...2

1. Work out the deviation (polygon - mean of neighbours)

top

- Value of unit (i) - mean(polygon - mean)
- Value of neighbour (j) - mean(polygon - mean) **bottom**
- Value of unit (i) - mean ^2 (polygon - mean)^2

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum (X_i - \bar{X})^2}$$

	x	value of polygon
	x hat	mean of values

Top part: Deviation= (value of unit (here it's I) - mean)*(value of neighbour (here it's J)-mean)

									Value of unit_unit - mean	
	W_ij	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	X_i	X_i - X ha	^2
	Zone 1	10.0489	6.8789	6.8789	-8.9711	-8.9711	-5.8011	7	3.17	10.0489
	Zone 2	6.8789	4.7089	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
	Zone 3	6.8789	4.7089	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
	Zone 4	-8.9711	-6.1411	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
	Zone 5	-8.9711	-6.1411	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
	Zone 6	-8.9711	-6.1411	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
unit	X_i	7	6	6	1	1	2	3.83		**43.4934**
unit-mean	(X_j - X ha	3.17	2.17	2.17	-2.83	-2.83	-1.83			

	W_ij	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	X_i	X_i - X ha	2
	Zone 1	10.0489	6.8789	6.8789	-8.9711	-8.9711	-5.8011	7	3.17	10.0489
	Zone 2	6.8789	4.7089	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
	Zone 3	6.8789	4.7089	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
	Zone 4	-8.9711	-6.1411	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
	Zone 5	-8.9711	-6.1411	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
	Zone 6	-8.9711	-6.1411	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
	X_i	7	6	6	1	1	2	3.83		**43.4934**
	(X_j - X ha	3.17	2.17	2.17	-2.83	-2.83	-1.83			

Example of Moran's I (global)...3

1. Times the deviation by the matrix (and sum the resulting matrix)

$$\frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum (X_i - \bar{X})^2}$$

f unit (here it's I) - mean)*(value of neighbour (here it's J)-mean)

Value of unit unit - mean

	Zone 3	Zone 4	Zone 5	Zone 6	X_i	X_i - X hat	^2
)	6.8789	-8.9711	-8.9711	-5.8011	7	3.17	10.0489
)	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
)	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
l	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
l	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
l	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
;	6	1	1	2	3.83		**43.4934**
'	2.17	-2.83	-2.83	-1.83			

Deviation * the matrix

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6
Zone 1	0	6.8789	6.8789	0	0	0
Zone 2	6.8789	0	4.7089	-6.1411	0	0
Zone 3	6.8789	4.7089	0	-6.1411	-6.1411	0
Zone 4	0	-6.1411	-6.1411	0	8.0089	5.1789
Zone 5	0	0	-6.1411	8.0089	0	5.1789
Zone 6	0	0	0	8.0089	8.0089	0

Sum of top part 42.4802

	Zone 3	Zone 4	Zone 5	Zone 6	X_i	X_i - X hat	2
)	6.8789	-8.9711	-8.9711	-5.8011	7	3.17	10.0489
)	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
)	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
l	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
l	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
l	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
;	6	1	1	2	3.83		**43.4934**
'	2.17	-2.83	-2.83	-1.83			

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6
Zone 1	0	3.43945	3.43945	0	0	0
Zone 2	2.29297	0	1.56963	-2.04703	0	0
Zone 3	1.71973	1.17723	0	-1.53528	-1.53528	0
Zone 4	0	-1.53528	-1.53528	0	2.00223	1.29473
Zone 5	0	0	-2.04703	2.66963	0	1.7263
Zone 6	0	0	0	4.00445	4.00445	0

Sum of top part 19.1051

Example of Moran's I (global)...4 (see next slide)

We now have:

- Number of units (N): 6
- Sum of weight matrix: (under N): 18 (binary) and 6 (row standardised)
- Sum of the deviation (times weight matrix - top of the right part): 42.46 (binary), 19.11 (row standardised)
- Sum of the deviation ^ 2 (bottom of the right part) = 43.39

Plug it in: (units (N)/ sum of weight matrix) * (weight matrix deviation / deviation squared)

- (6/18) * (42.48/43.49) = 0.33
- (6/6) * (19.11/43.49) = 0.44

Example of Moran's I (global)...4

$$\frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum (X_i - \bar{X})^2}$$

f unit (here it's I) - mean)*(value of neighbour (here it's J)-mean)

Value of unit unit - mean

	Zone 3	Zone 4	Zone 5	Zone 6	X_i	X_i - X ha	^2
)	6.8789	-8.9711	-8.9711	-5.8011	7	3.17	10.0489
)	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
)	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
L	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
L	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
L	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
;	6	1	1	2	3.83		**43.4934**
'	2.17	-2.83	-2.83	-1.83			

Deviation * the matrix

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6
Zone 1	0	6.8789	6.8789	0	0	0
Zone 2	6.8789	0	4.7089	-6.1411	0	0
Zone 3	6.8789	4.7089	0	-6.1411	-6.1411	0
Zone 4	0	-6.1411	-6.1411	0	8.0089	5.1789
Zone 5	0	0	-6.1411	8.0089	0	5.1789
Zone 6	0	0	0	8.0089	8.0089	0

Sum of top part 42.4802

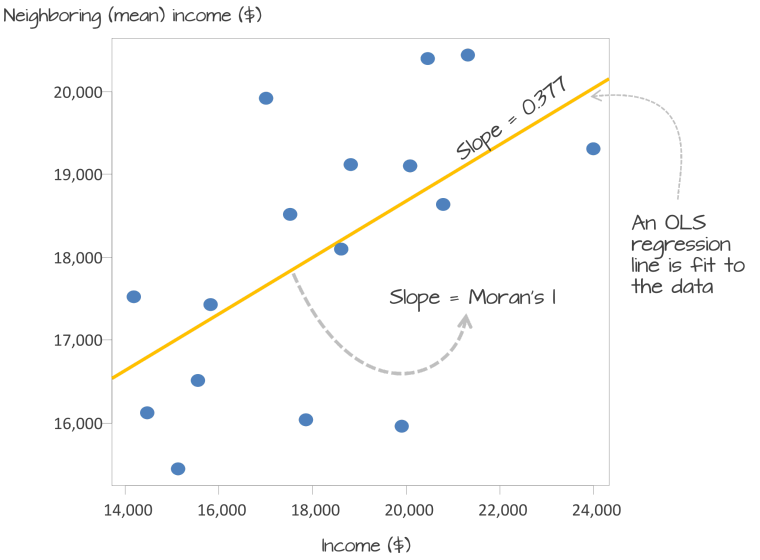
	Zone 3	Zone 4	Zone 5	Zone 6	X_i	X_i - X hat	2
)	6.8789	-8.9711	-8.9711	-5.8011	7	3.17	10.0489
)	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
)	4.7089	-6.1411	-6.1411	-3.9711	6	2.17	4.7089
L	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
L	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
L	-6.1411	8.0089	8.0089	5.1789	1	-2.83	8.0089
;	6	1	1	2	3.83		**43.4934**
'	2.17	-2.83	-2.83	-1.83			

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6
Zone 1	0	3.43945	3.43945	0	0	0
Zone 2	2.29297	0	1.56963	-2.04703	0	0
Zone 3	1.71973	1.17723	0	-1.53528	-1.53528	0
Zone 4	0	-1.53528	-1.53528	0	2.00223	1.29473
Zone 5	0	0	-2.04703	2.66963	0	1.7263
Zone 6	0	0	0	4.00445	4.00445	0

Sum of top part 19.1051

Another way to think about this..Moran plot

- In the Moran plot, we plot the value of spatial unit (against the average neighbouring values)
    - They might be row standardised
- **Note**, this idea if comparing our value to neighbouring values is called a **spatial lag**
- This shows us the relationship between the variables (x: raw value, y **average** standardised values from neighbors)
    - If we used binary then this is the sum
    - row standardised is a local average



Moran plot. Source:Gimond, 2022

The slope is the Moran's I value (or coefficient here)

Plugging in the numbers

....head to the excel document for a Moran's I example with the formula

Other measures of spatial autocorrelation

- Geary's C
  - This tells us whether **similar values or dissimilar values are clustering**
  - Geary's C falls between 0 and 2; 1 means no spatial autocorrelation, <1 - positive spatial autocorrelation or similar values clustering, >1 - negative spatial autocorreation or dissimilar values clustering) which shows that similar values are clustering
- Getis Ord General G
  - This tells us whether **high or low values are clustering**.
  - If G > Expected = High values clustering; if G < expected = low values clustering
- Moran's I
  - Tells us whether **we have clustered values (close to 1) or dispersed values (close to -1)**
  - Nothing about if the values are high or low

Other measures of spatial autocorrelation

- Geary's C and Moran's I are almost inversely related to each other...but they are not identical
- Moran's I = global comparison
  - standardizing the spatial autocovariance by the variance of the data
- Geary's C = sensitive to local spatial autocorrelation
  - the sum of the squared differences between pairs of data values as its measure of covariation

Formula

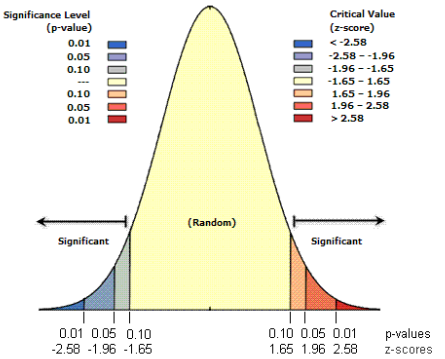
$$I = \frac{N}{S_0} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j)(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, j \neq i \quad [1]$$

$$c = \frac{(N-1)}{2S_0} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j)(x_i - x_j)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, j \neq i \quad [2]$$

Moran I and Geary C. Source:NKU, 2022

Local versions (LISA) - Anselin Moran's I

- Global Moran's I compares each spatial unit to the neighbours and then gives an average of all the differences identified
- Local Moran's I `localmoran(., Lward.lw)` gives a value for each spatial unit in relation to neigbhours and ...
    - Moran's I value
    - Expected Moran's I
    - Variance (of neighbors)
    - Z score - needs the expected and variance
    - p-value
- z score here allows us to state if our value is significantly different than expected at this location considering the neighours.



What is a z score. Source:ArcMap

- Z scores = standard deviations
- High or low = unlikely that it is completely spatially random
- We reject the null hypothesis (complete spatial randomness)
- Middle of the normal distribution is the expected outcome

Global vs Local Spatial autocorrelation. Source:story maps

Local versions (LISA) - Anselin Moran's I

Null hypothesis of complete spatial randomness

> "A null hypothesis is a type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations"

- We assume that there is complete spatial randomness as default
- We are testing against that assumption
- The assumption is that the value will be within < -1.65 or > +1.65 standard deviations from the mean (or whatever you define - see later slide with z score table)
- If it isn't then we are **far enough** away from the mean to state that this isn't random, based on the neighboring values.
- Compared to the surrounding values, this one is different

Local versions (LISA) - Anselin Moran's I

Local Moran's I explained

- Difference between the value and mean (or neighbourhood) * sum of differences between neighbours and mean

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1}^{n} w_{i,j}(x_j - \bar{X})$$

GLOBAL VS LOCAL SPATIAL AUTOCORRELATION. Source:Geog 418/518

- Where the difference between the value of the unit (i) and the mean is divided by the standard deviation...
  - Standard deviation = how much the values in the neighbourhood vary from mean of neighbourhood

$$S_i^2 = \frac{\sum_{j=1}^{n}(x_j - \bar{X})^2}{n-1}$$

GLOBAL VS LOCAL SPATIAL AUTOCORRELATION. Source:Geog 418/518

Local versions (LISA) - Anselin Moran's I

Local Moran's I explained

- z-test = compare our local value of i to the neighbourhood if it were random...

$$Z_{I_i} = \frac{I_i - \mathrm{E}[I_i]}{\sqrt{V[I_i]}}$$

GLOBAL VS LOCAL SPATIAL AUTOCORRELATION. Source:Geog 418/518

- Where the expected Moran's I value at location is computed from a formula
- And the variance between of Moran's I also computed...

Local versions (LISA) - Anselin Moran's I

Z score shows us...

- z-test = compare our local value of i to the neighbourhood if it were random...
  - (Local i - minus sum of values in neighborhood / total number -1) / square root of variance

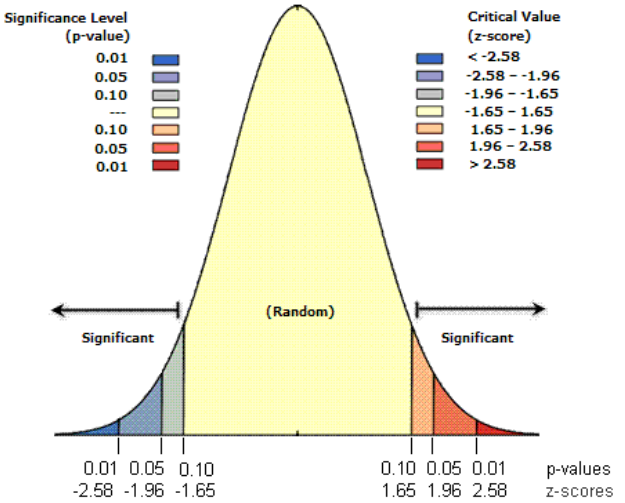
$$Z_{I_i} = \frac{I_i - \mathrm{E}[I_i]}{\sqrt{V[I_i]}}$$

GLOBAL VS LOCAL SPATIAL AUTOCORRELATION. Source:Geog 418/518

See: GLOBAL VS LOCAL SPATIAL AUTOCORRELATION for formulas

Local versions (LISA) - Anselin Moran's I

- Once we have the z score we compare it to the confidence levels to see it was by complete spatial randomness
- If it does we expect it to be in the middle of the distribution...Null hypothesis...
- If it is within the tails then it is not completely random ...reject the null hypthoesis
- The Moran's I value (or +ve /-ve Z score) will tell us if it is clustering (+ve) or dispersion (-ve)...



What is a z score. Source:ArcMap

Local versions (LISA) - Anselin Moran's I

- z-score interpretation is as follows:

z-score (Standard Deviations)	p-value (Probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

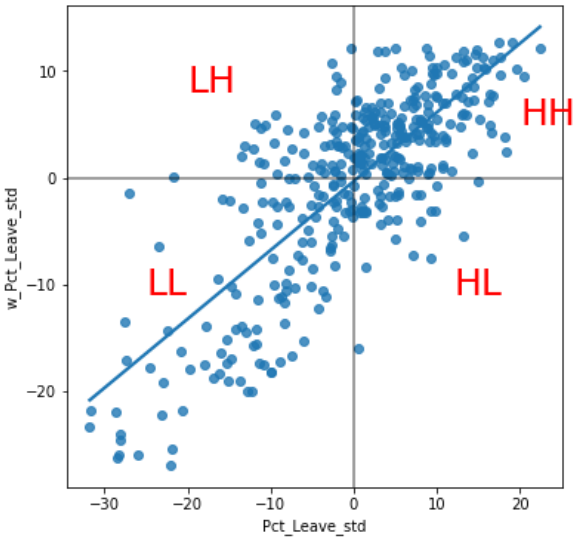
What is a z score. Source:ArcMap

We interpret this as...

> `>2.58` or `<-2.58` standard deviations away from the mean are significant at the 99% level...this means there is a <1% chance that autocorrelation not present
>
> `>1.96 - <2.58` or`<-1.96 to >-2.58` are significant at the 95% level (<5% change that autocorrelation not present)

Local versions (LISA) - Anselin Moran's I

- **Core idea** is to identify areas that are more similar (or dissimilar) to surroundings based on what we expect by chance!
- In this case Prof Dani Arribas-Bel has standardised the values (subtract average and divided by standard deviation) for brexit votes
- This gives us those with above average leave voting and below average leave voting



Moran plot. Source:Arribas-Bel, 2022

- high values surrounded by high values (HH)
- low values nearby other low values (LL)
- low values among high values (LH)
- high values among low values (HL)

Local versions (LISA) - Anselin Moran's I

Combine this idea with significance (either p or z score)

I have done this using `case_when()`!

```
signif <- 0.1


# centers the variable of interest around its mean
points_sf_joined2 <- points_sf_joined %>%
  mutate(mean_density = density- mean(density))%>%
  mutate(mean_density = as.vector(mean_density))%>%
  mutate(mean_densityI= density_I - mean(density_I))%>%
  mutate(quadrant = case_when(mean_density>0 & mean_densityI >0 ~ 4,
         mean_density<0 & mean_densityI <0 ~ 1,
         mean_density<0 & mean_densityI >0 ~ 2,
         mean_density>0 & mean_densityI <0 ~ 3))%>%
  mutate(quadrant=case_when(p > signif ~ 0, TRUE ~ quadrant))
```

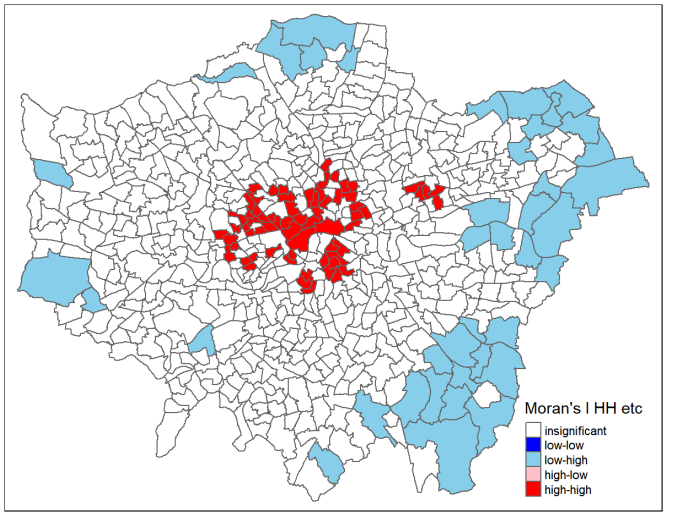
See: https://rpubs.com/AndrewMacLachlan/870348

Local versions (LISA) - Anselin Moran's I

Combine these ideas - significant and surrounded by like values....

- high values surrounded by high values (HH)
- low values nearby other low values (LL)
- low values among high values (LH)
- high values among low values (HL)

Pharmacy density in London.



Source:Andy MacLachlan

Local versions - Anselins Moran's I

Remember what the values mean

Moran's I tells us whether we have clustered values (close to 1) or dispersed values (close to -1)

for local this is in relation to its neighbours...

for global it's all values as an average

Local versions - Getis-Ord Gi*

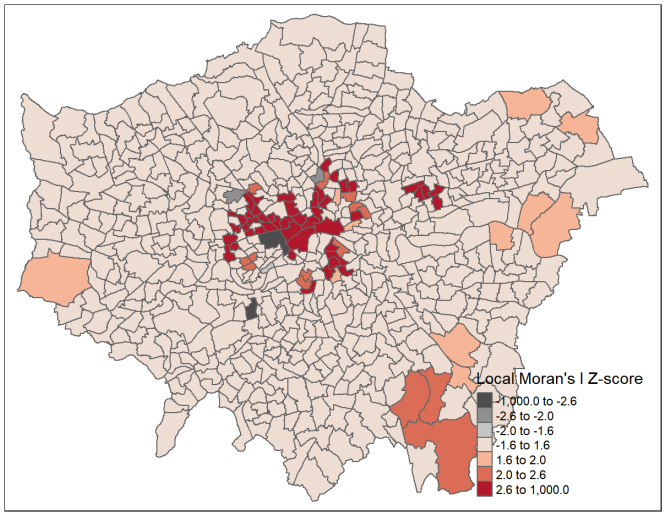
- Pronounced G-i-star
- Similar concept but **just returns a z score**
    - standardised value relating to whether high values or low values are clustering together
    - However....this is the local sum comapred to the **sum of all features**
    - A high value = sum of values within a neighborhood of a given radius or configuration is high relative to the **global average**
- More intense clustering of high values = hot spot
- More intense clustering of low values = cold spot

```
localG(., Lward.lw)
```

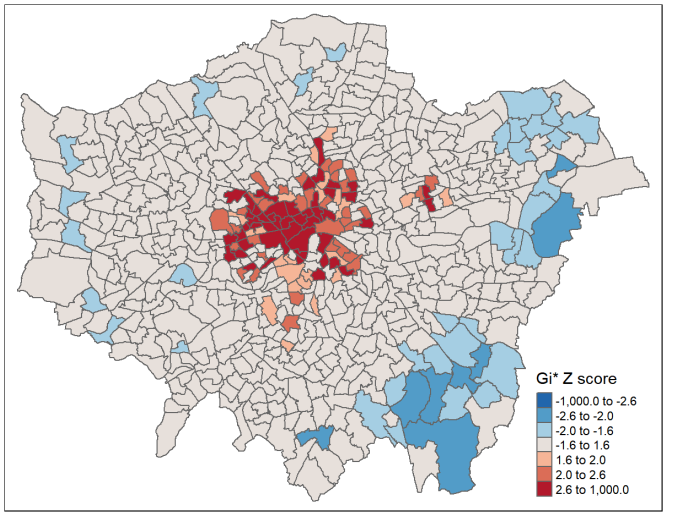
Moran's I vs Getis-Ord Gi*

Pharmacy denisty in London

- Moran's I



- Getis Ord Gi*



Moran's I vs Getis-Ord Gi* notes

- Moran's I is a measure of the degree to which the value at a target site is similar to values at adjacent sites. Moran's I is large and positive when the value for a given target (or for all locations in the global case) is similar to adjacent values and negative when the value at a target is dissimilar to adjacent values.

- Getis Ord G∗i identifies areas where high or low values cluster in space. It is high where the sum of values within a neighborhood of a given radius or configuration is high relative to the global average and negative where the sum of values within a neighborhood are small relative to the global average and approaches 0 at intermediate values.

Limitations

- Calculating the neighbours is computationally intensive. Creates an n^2 matrix.
- So 10 spatial units is manageable, but 1000 requires 1,000,000 comparisons, and 10,000 100,000,000 etc.
- This is very important as defining spatial weights matrices is crucial for running other types of analysis – spatially-lagged regression or Geographical Weighted Regression (which we will come onto in subsequent weeks)
- Global measures can mask local trends (although local versions deal with this to an extent)
- Indices of spatial association can be a good summary of the system, but it's never possible to capture all complex spatial interactions in a single figure
- **DO NOT COMPARE IF THEY HAVE DIFFERENT ZONES**

The next question...

We know where the clusters are but why are they there....or can we explain what is happening...

Summary

Points

- Are these points distributed in a random way or is there some sort of pattern (uniform or clustered)?
- Point Pattern Analysis techniques all compare observed distributions of points to an expected model based on the Poisson distribution – with varying degrees of sophistication

Spatially continuous observations (e.g. values of polygons)

- How (dis)similar are our values assigned to geographic units across geographic space
- Analysis of the spatial autocorrelation of continuous variables over space allows us to assess if similar values cluster in space
- How we define neighbours is crucial in spatial autocorrelation analysis

Final task

Consider the tools for point pattern anslysis and spatial autocorrelcation and write down what each can show...e.g...

Moran's I shows if neighbouring places have more similar (or dissimilar) values when compared with their difference to the average in the system

Local Moran's I shows....?