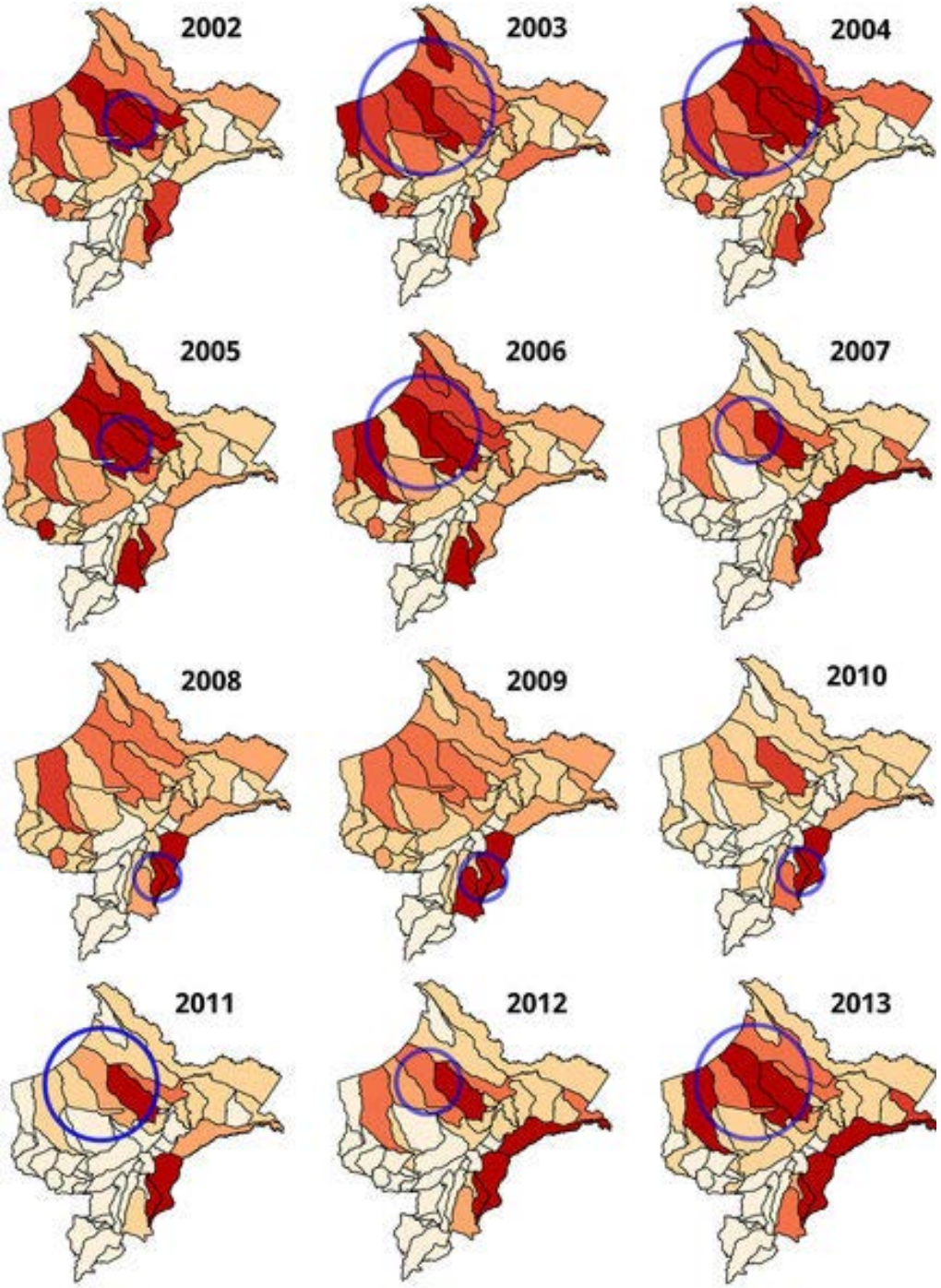



# Analyzing spatial clustering

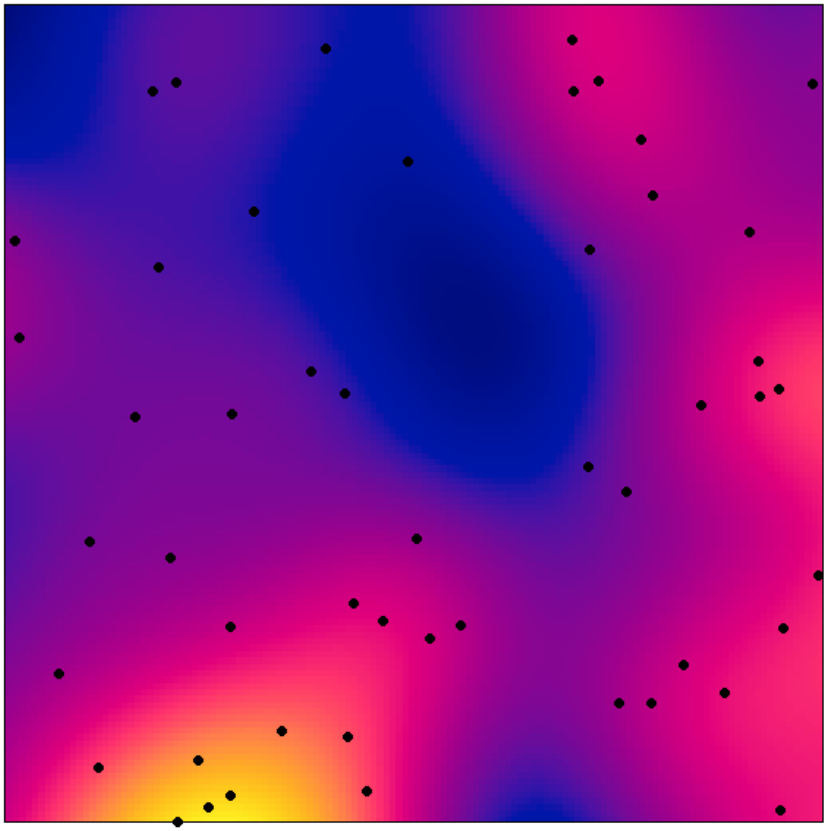
Jennifer Smith

# What is clustering?

“Spatial autocorrelation”

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

Tobler’s First law of Geography



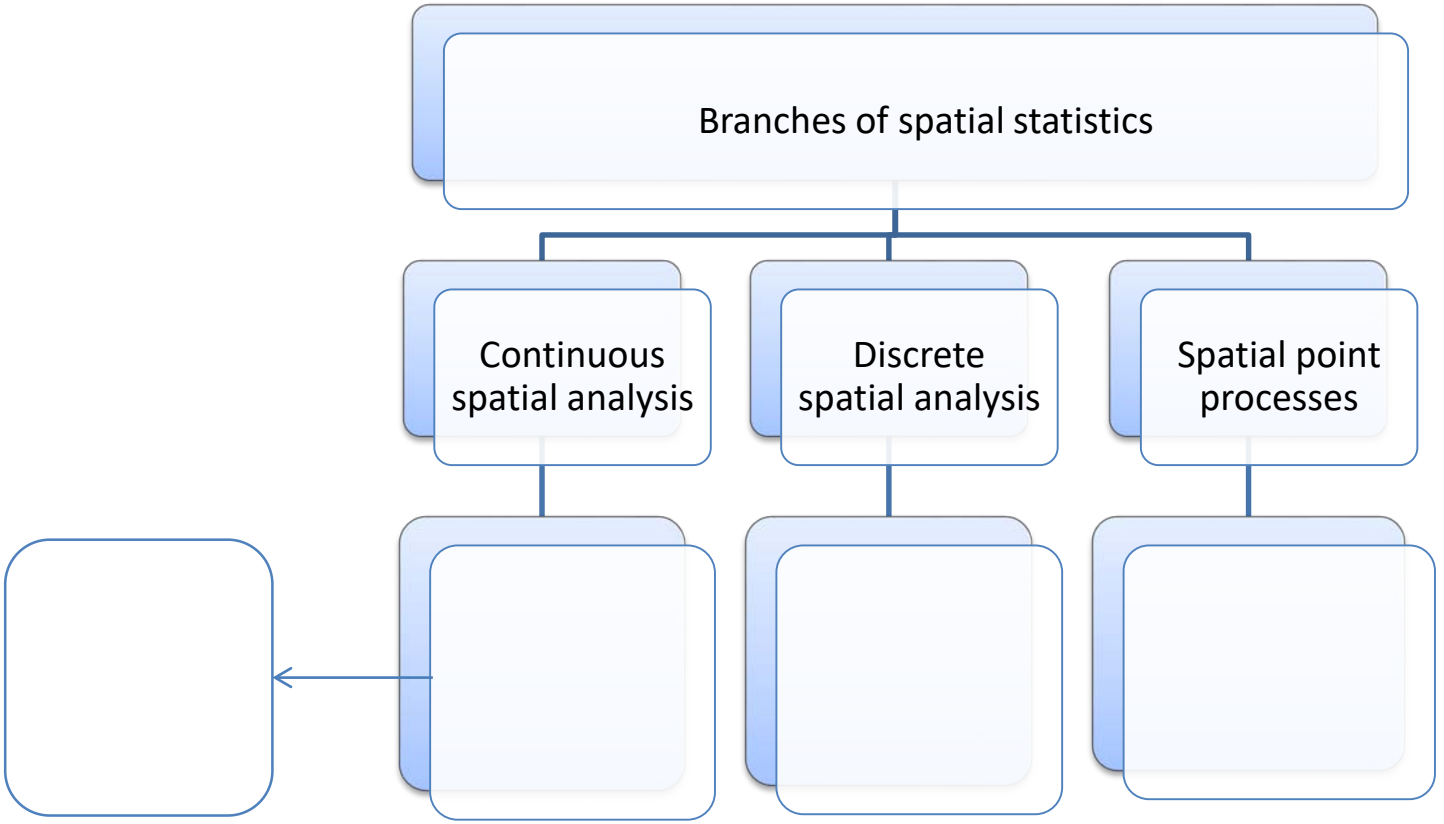
# Mechanisms for clustering

- Infectious diseases (direct transmission/ vector distribution)
- Shared risk factors (i.e. socioeconomic status)
- Health hazards (i.e. localized pollution sources, etc)
- Data anomalies with spatial pattern (i.e. reporting bias)

# Ways to define clustering

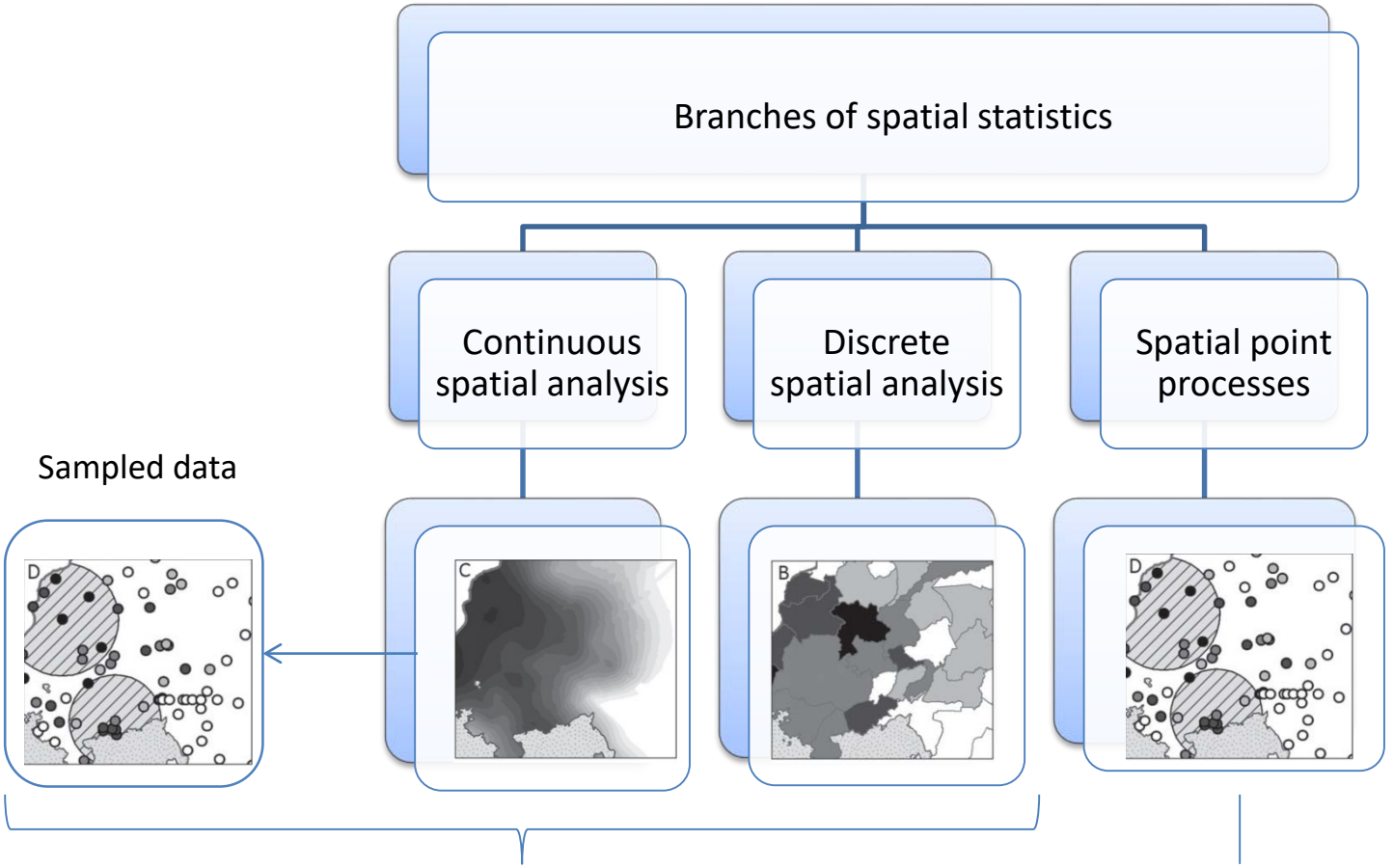
- Global vs local clustering methods
  - i.e. *clustering* vs *cluster detection*
- Raw vs residual
  - i.e. corrected for known risk factors
- Scale
  - Clustering likely to happen over multiple scales
  - Clustering over larger spatial scales (and these tests of association) may mask more local clusters

# Data types




*Pullan et al (2012)*

# Data types

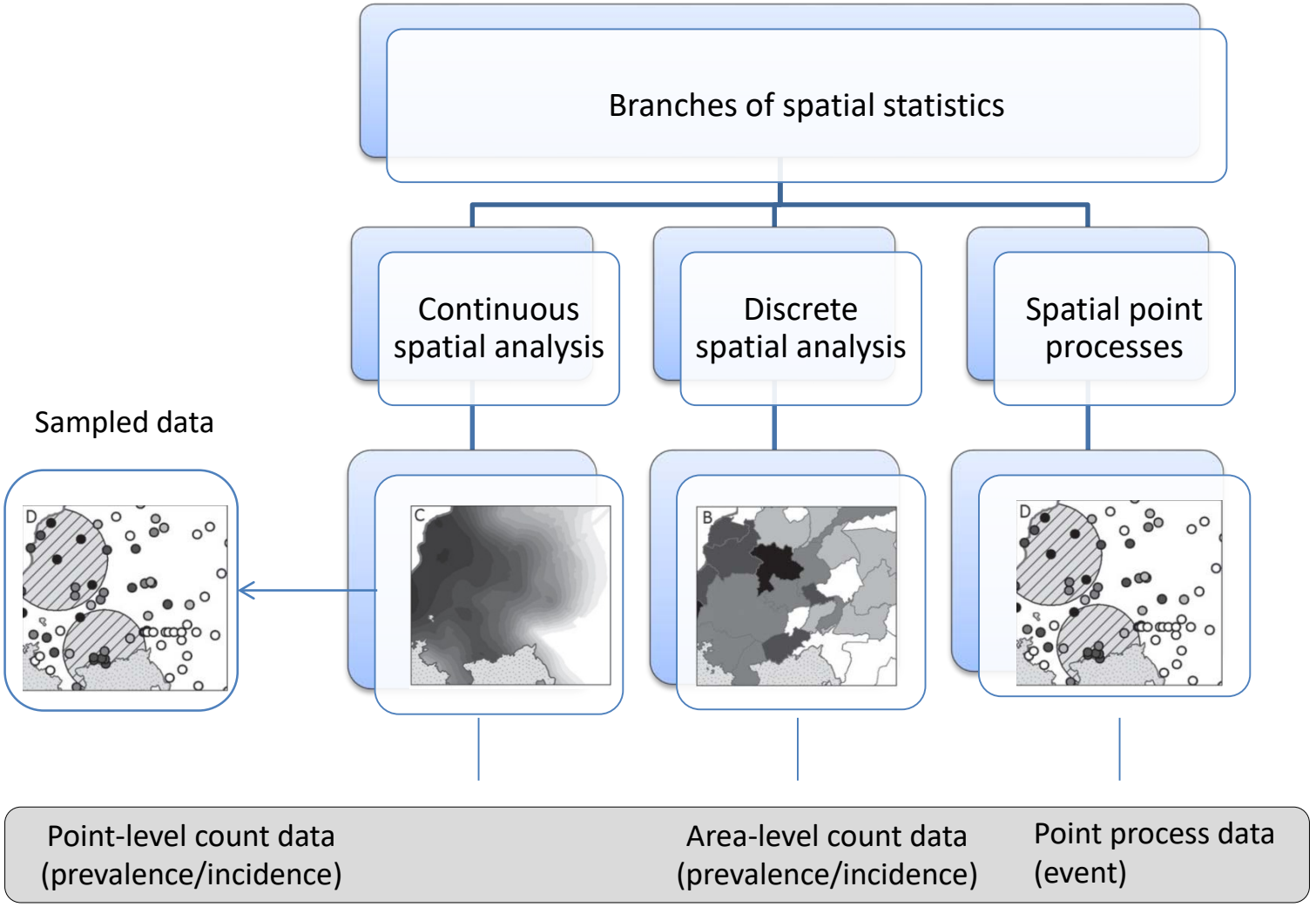



Rely upon spatially sampled measurement data
Global (i.e. model overall spatial autocorrelation in data)

Physical location of events in region
Global **propensity to cluster** or **local spatial clusters** of infection

*Pullan et al (2012)*

# Data types


Branches of spatial statistics
Continuous spatial analysis
Discrete spatial analysis
Spatial point processes
Sampled data
Point-level count data (prevalence/incidence)
Area-level count data (prevalence/incidence)
Point process data (event)

# Variation amongst attribute values

(aka "**first-order spatial effect**")
Variation in the mean

(aka "**second-order spatial effect**"
Residuals


Large scale trends

- Geostatistical: mean of the outcome
- Point process: mean intensity

Local dependence of spatial process & smaller scale interactions between neighbors

- Geostatistical: autocorrelation
- Point process: clusters

# First-order characteristics (macro)

- Describe the way in which the expected value (mean or average) of the spatial pattern varies across space (i.e., the density of the spatial point pattern)
- May be constant or spatially varying, but assumes that individual event locations are independent of one another (i.e. no spatial dependency).

<u>i.e. where do patterns appear to differ?</u>

# Second-order properties (meso-/micro-)

- Inform on the interrelationship between events or spatial dependence.
- Measures by a dependence measure – e.g. nearest neighbor analysis, Ripley's k-function
- Insight into global aspects of the point pattern

<u>i.e. are there general patterns of clustering with respect to CSR or another pattern?</u>

# Terminology and overlap

Related to scale and sampling methods

| Point-level data | Area data | Raster data | Point-process data |
|---|---|---|---|
| | Areal data | Continuous data | Event data |
| | Aggregate data | Grid data | Point data |

| Geostatistical data | Point-process data |
|---|---|
| Global | Local |
| Continuous variation | Point pattern |
| Clustering | Cluster detection |
| Spatial autocorrelation | Cluster analysis |

# Point-level vs Point-process data

| | Point-level data | Point process data |
|---|---|---|
| Measurement | Could in principle be measured anywhere, but typically come at a limited number of observation locations. | Corresponds to case-ascertainment. |
| Locations | Sampled locations usually based on “representativeness” or random sampling. | Usually event locations (i.e. incident case) over a given time frame and study region. |
| Analytic objectives | Pattern of observation locations itself usually not of primary interest – rather to infer aspects of the variable that have not been measured (i.e. prevalence over larger study area) | Determine whether the pattern of event locations contains clusters of events with increased risk of disease. |
| Outcome | Prevalence/Incidence (count) | Density |

# Aims of cluster analysis

- Are observed patterns ‘random’?
  - Point-processes define clustering as a departure from this randomness
- What is the propensity for spatial clustering or locations of individual clusters (places with excess cases)?
  - Surveillance systems
- Where are areas with higher than expected risk?
  - Target interventions to high risk areas; detect potential outbreaks
- Over what scale does clustering occur?
  - Determine the optimal scale of interventions

# Main types of tests for spatial clusters

**Global statistics**

- propensity for clustering anywhere in the study area, e.g. Ripley's *K*-function, Cuzick & Edwards test for point data and Moran's I for area data

**Local statistics**

- cluster(s) in specific locations, e.g. Spatial scan statistic for point data and Getis and Ord's local G*i*(d) statistic for area data

**Focused statistics**

- cluster(s) around a specific point source, e.g. Diggle's method, Stone's test, Bithell's test

# Global and Local tests for spatial clustering: point-level and area data

# Data

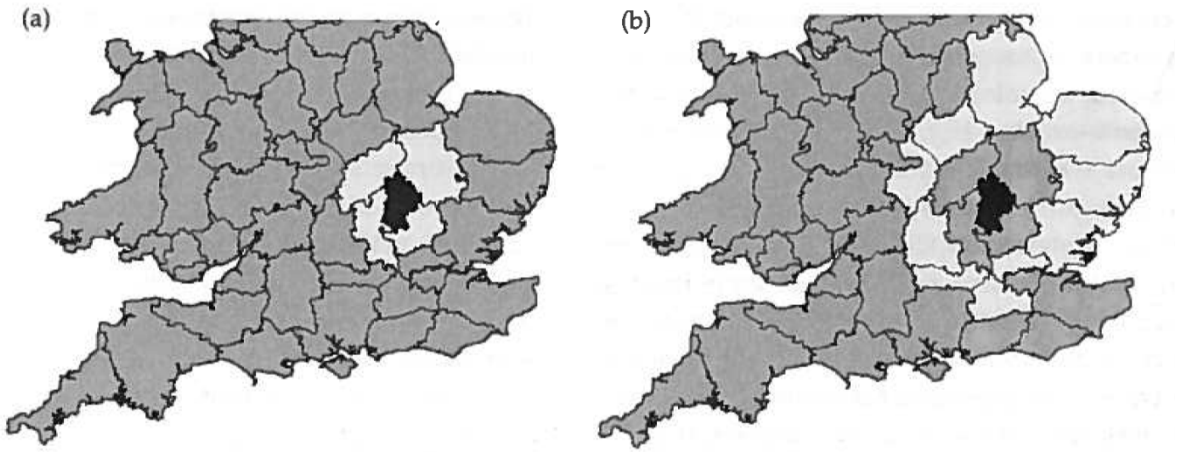
**Point-level & area-level data:** reported counts of incident (newly diagnosed) or prevalent (existing) cases residing in particular areas within the study area

- Point prevalence of malaria in Burkina Faso
- Point prevalence of malaria in The Gambia
- Area level incidence of lip cancer in Scotland

# Autocorrelation statistics

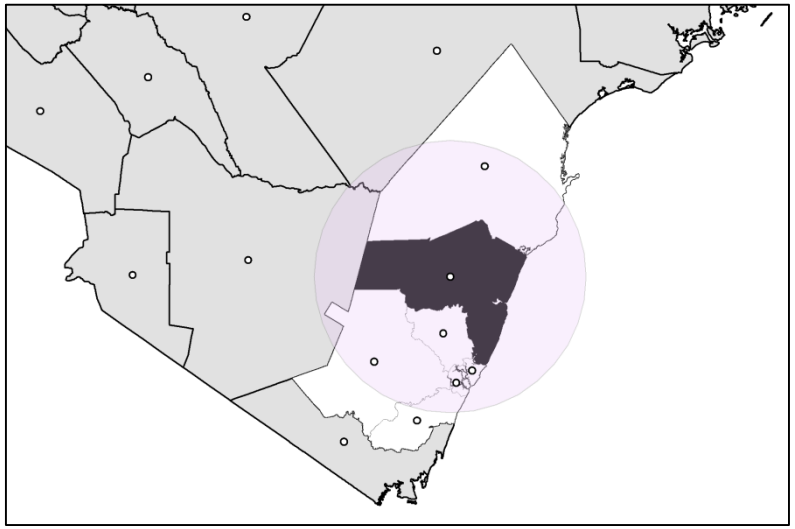
- Provide an estimate of the degree of spatial similarity observed among neighboring values of an attribute (i.e. disease prevalence) over a study area.
- Requires:
  1. Definition of neighbors (distance-based or adjacency)
  2. Weights matrix (those close in space given greater weight)

# Defining neighbors




Adjacency

a) first-order
b) Second-order



Distance-based

NB: may choose distance that allows each point/polygon at least one neighbor OR define spatial lags

# Defining spatial weights

- In order to calculate summary statistics, you have to assign spatial weights to each relationship
  - Binary: used with fixed distance, K nearest neighbors, and contiguity spatial relationships. Neighbors =1, non-neighbors=0
  - Variable: For inverse distance or inverse time spatial relationships, weights fall into a range from 0 to 1 with nearby neighbors getting larger weights than neighbors farther away.
- Rule of thumb – stick to binary representation unless you know about the spatial process.
- Default in R is “W” where the weights are standardized to sum to unit (aka row standardization).

# Moran's I statistic

- Intended for use with continuous data (although it can also be used with count data – issues with population distribution)

$$I = \frac{n \sum_i \sum_j W_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\left(\sum_i \sum_j W_{ij}\right) \sum (Z_i - \bar{Z})^2}$$

Weights matrix

Deviation of feature attribute from mean

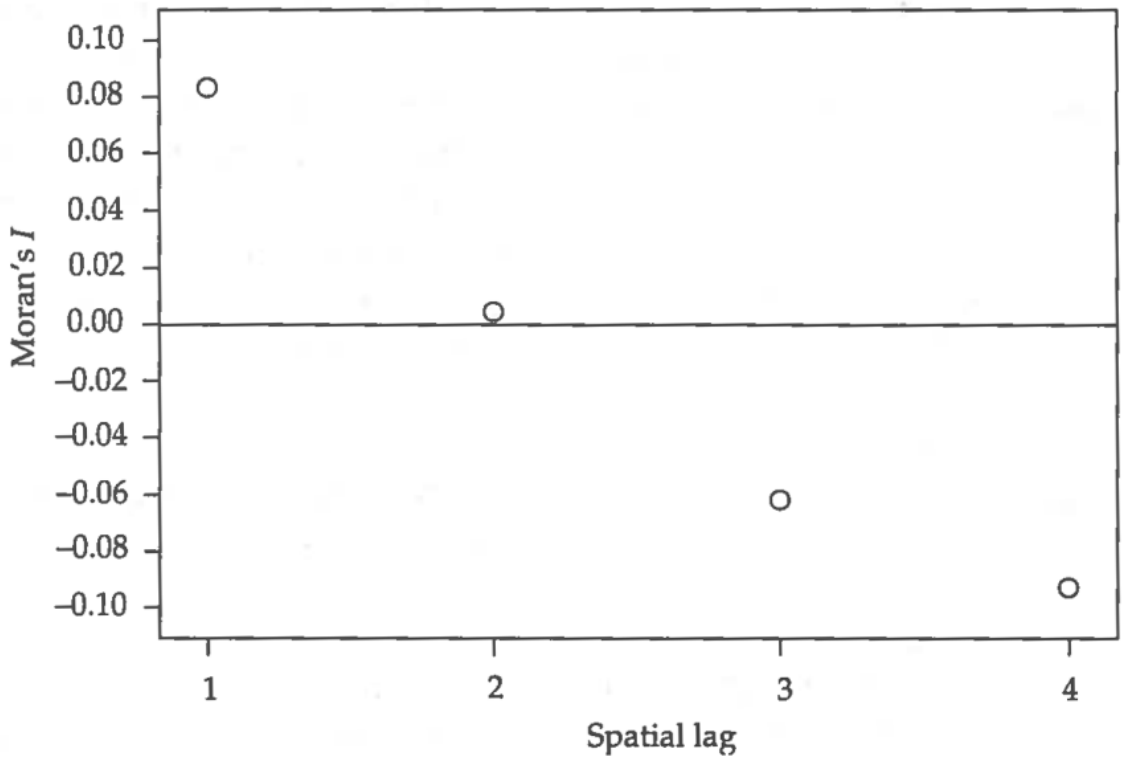
Total number of features

Aggregate of all the spatial weights

- Weights matrix (W) is used to define the spatial relationship (regions close in space given more weight calculating the statistic)
- Generally between +1 and -1; value of 0 indicates no clustering
- If there is clustering (positive spatial autocorrelation) then areas close together will have similar responses and the term $(Z_i -$

# Correlogram of Moran's I

- Series of estimates of Moran's I evaluated at increasing distances and plotted against one another – can help to determine where spatial autocorrelation is maximized.



# Hypothesis testing

- Use simulation (Monte Carlo randomization) to assess the significance of clustering
  - i.e. does the observed spatial pattern differ significantly from the null hypothesis of complete spatial randomness
- Calculate the test statistic using the observed data and then re-calculating it using a specified number of simulated data sets (or permutations).
  - Expected distribution of the test statistic under the null hypothesis
  - Calculate the likelihood of obtaining the value for the test statistic derived from the observed data and expressed as the p-value = the proportion of the test statistic value (from the simulated data) that are greater than the value of the test statistic from the observed data.

# Local Indicators of Spatial Association (LISAs)

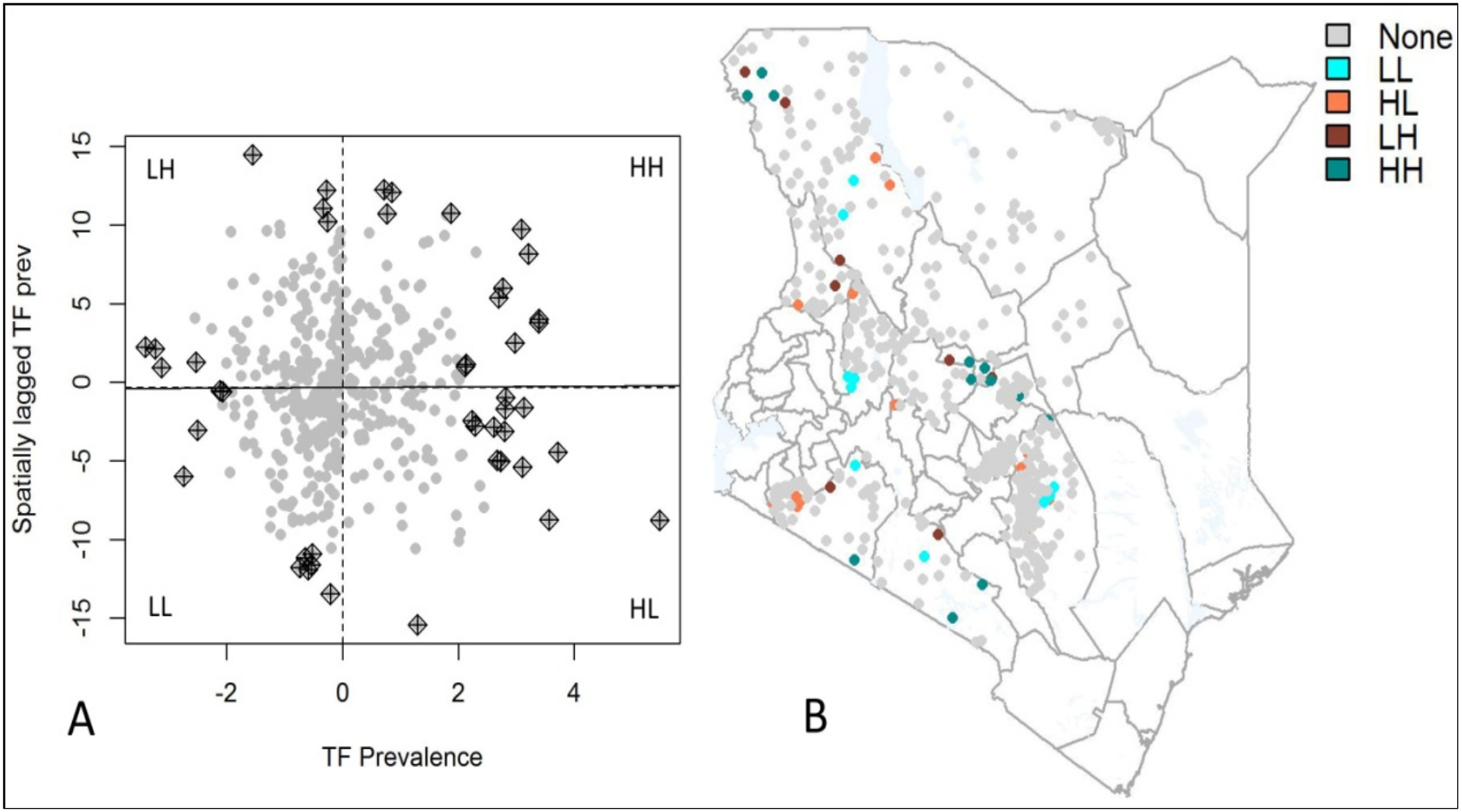
- Global measures assume that the spatial process is stationary – if this is not true than tests can obscure local effects.
- Particularly an issue with larger datasets because they summarize large amounts of potentially dissimilar information
- LISAs scan the whole dataset but only measure dependence in limited portions of the study area
- Detection of disease clusters……

# Local Moran's

- Decomposes Moran's I statistic into contributions for each area.

$$I_i = Z_i \sum_{j, j \neq i}^{n} w_{ij} Z_j$$

- Detects clusters of either similar or dissimilar disease frequency values around a given observation
- Sum of the LISAs is proportional to the global Moran's I statistic.


LH
HH
LL
HL
Spatially lagged TF prev
15
10
5
0
-5
-10
-15
-2
0
2
4
TF Prevalence
A
B
None
LL
HL
LH
HH

# Example and in-class exercises

- Import point malaria prevalence data from Burkina Faso
- Visualize data
- Calculate global Moran's I and a correlogram
- Calculate local Moran's I

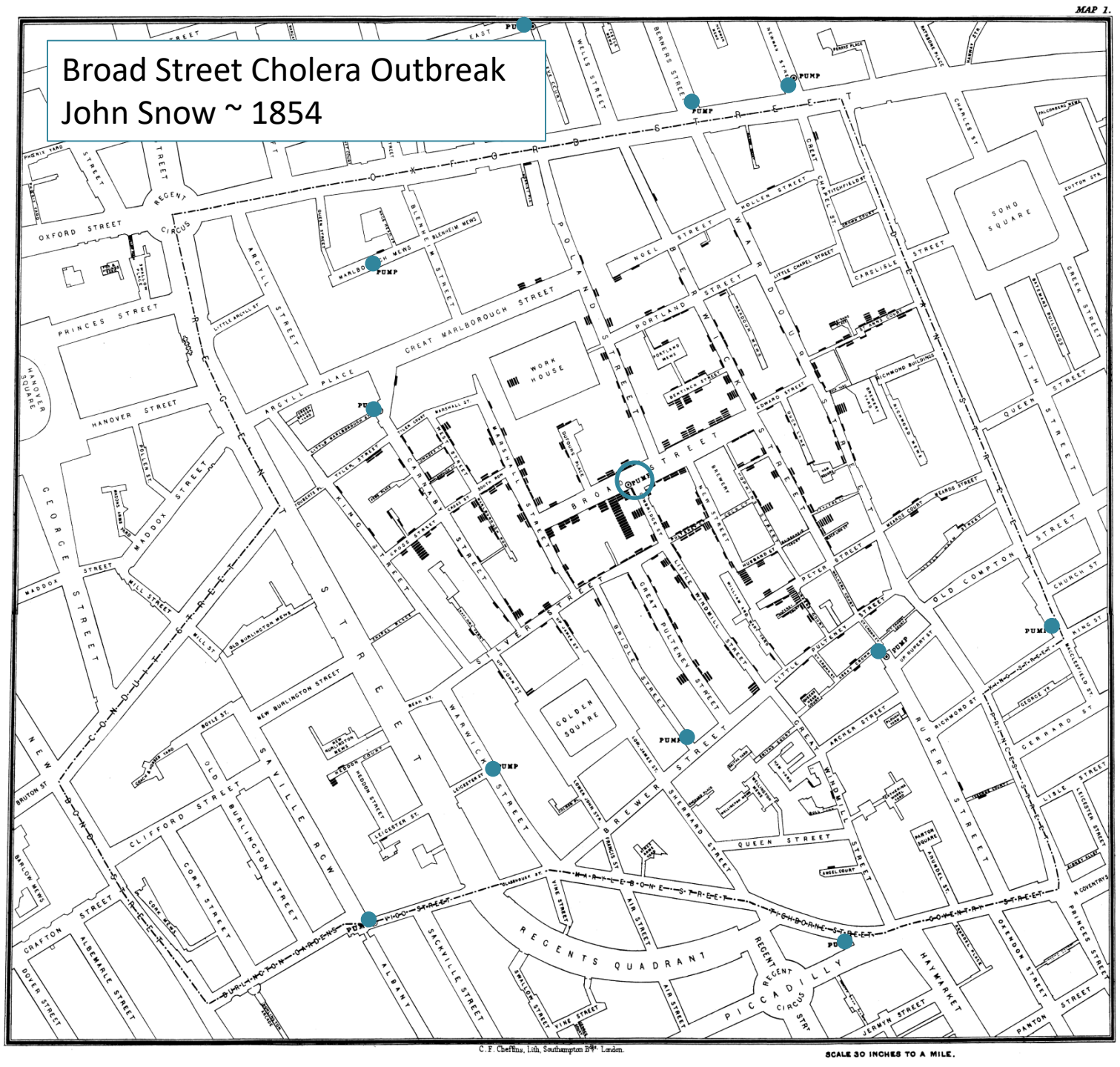
In class exercises: reproduce with the point level data from the Gambia and area-level data on Scottish lip cancer

# Introduction to point processes

# Data

**Case-control point data:** point locations for each of a set of cases reported and a collection of non-cases (controls). “Marked point data”

- Malaria cases and population controls from Namibia



C.F. Cheffins, Lith, Southampton Bldgs. London.

SCALE 30 INCHES TO A MILE.

# Properties of spatial point patterns

- A point pattern is a set of locations in a defined study region at which events of interest have been recorded

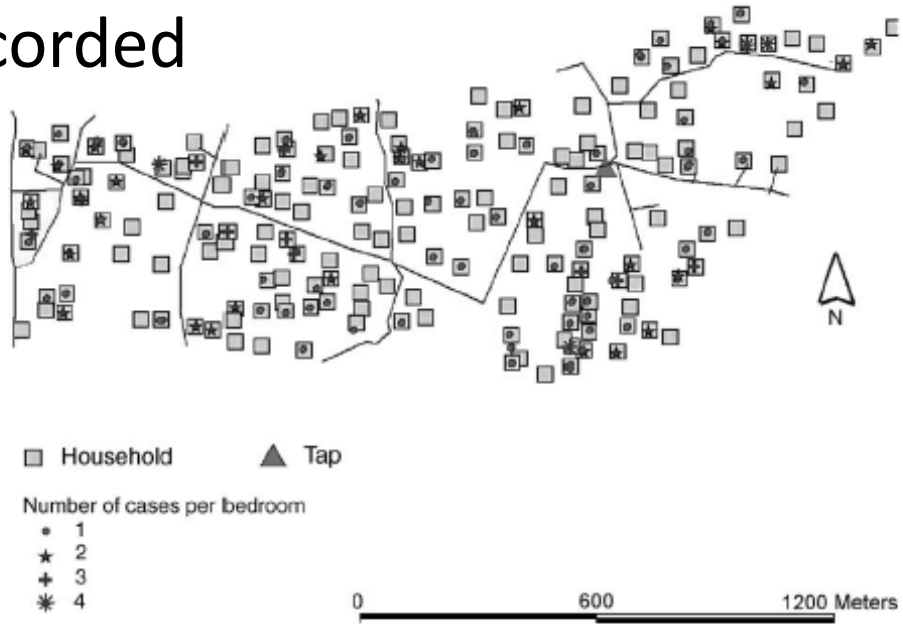



**Figure 3** Distribution of cases of active trachoma by bedroom in Kahe Mpya, Tanzania: location of bedrooms within households with active trachoma.

Polack et al (2005)

**Key terms**

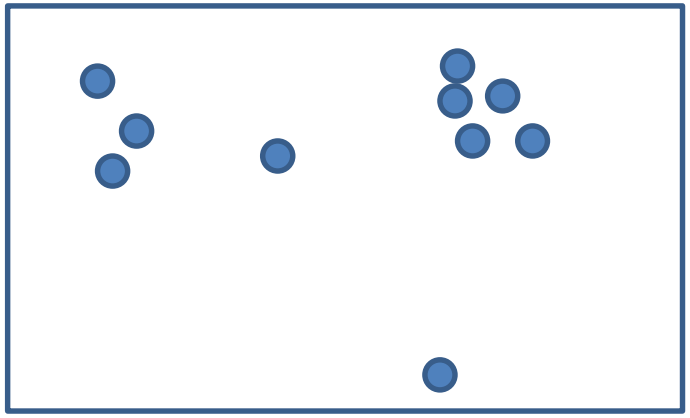
***Event*** – an occurrence of interest (e.g. incident case of disease)

***Point*** – any location in the study area where an event could occur

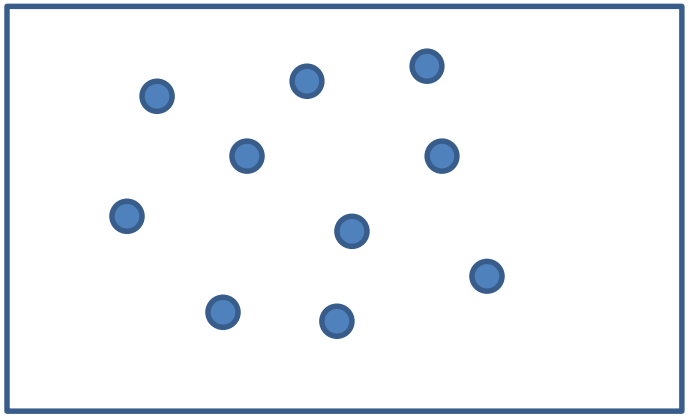
***(Event) Location*** – specific location where an event did occur

***Data set*** – collection of observed event locations and a spatial domain of interest

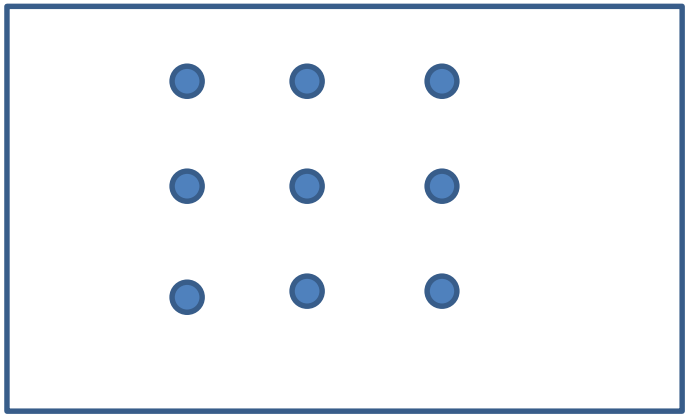
# General forms of point patterns



**Clustered –** some points clustered together, other areas contain no points



**Random** – a given point is equally likely to occur at any location and independent of other points



**Regular** – each point is as far from neighbour as possible

# What is a point process?

- An observed point pattern that can be described by a spatial stochastic process (probabilistic model)
- Any observed spatial point pattern can be considered an outcome (or a realisation) of a spatial stochastic process – i.e. a data set resulting from a particular model
- Typically a Poisson process

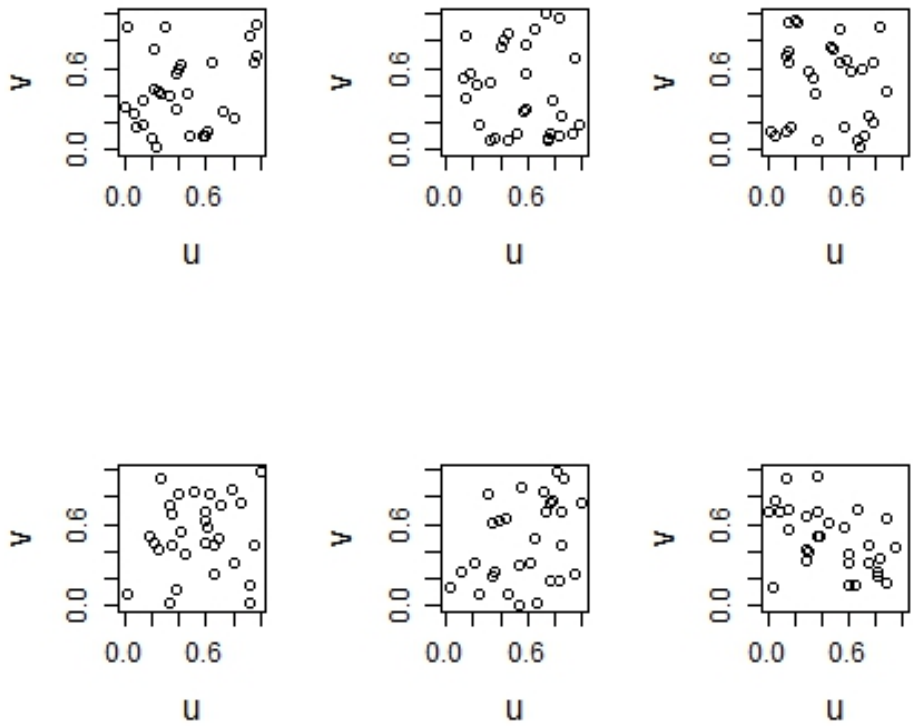
# Components of a point process

- First-order (mean) properties
  - Describes the expected density of points through space.
  - Measured by an intensity measure, e.g. quadrat estimation, kernel estimation (next lecture)
- Second-order properties
  - Informs on the interrelationship between events of the process
  - Similar to variance and covariance
  - Summarize spatial dependence between events over different spatial scales

# Complete spatial randomness (CSR)

- Homogeneous Poisson process (HPP)
- Simplest theoretical model for a spatial point pattern (stationary; independence of points)
- Events are distributed independently according to a uniform probability distribution over the region
- Used as a benchmark for hypothesis testing – not useful if departure is obvious due to a priori heterogeneity (e.g. underlying population distribution)

# CSR II



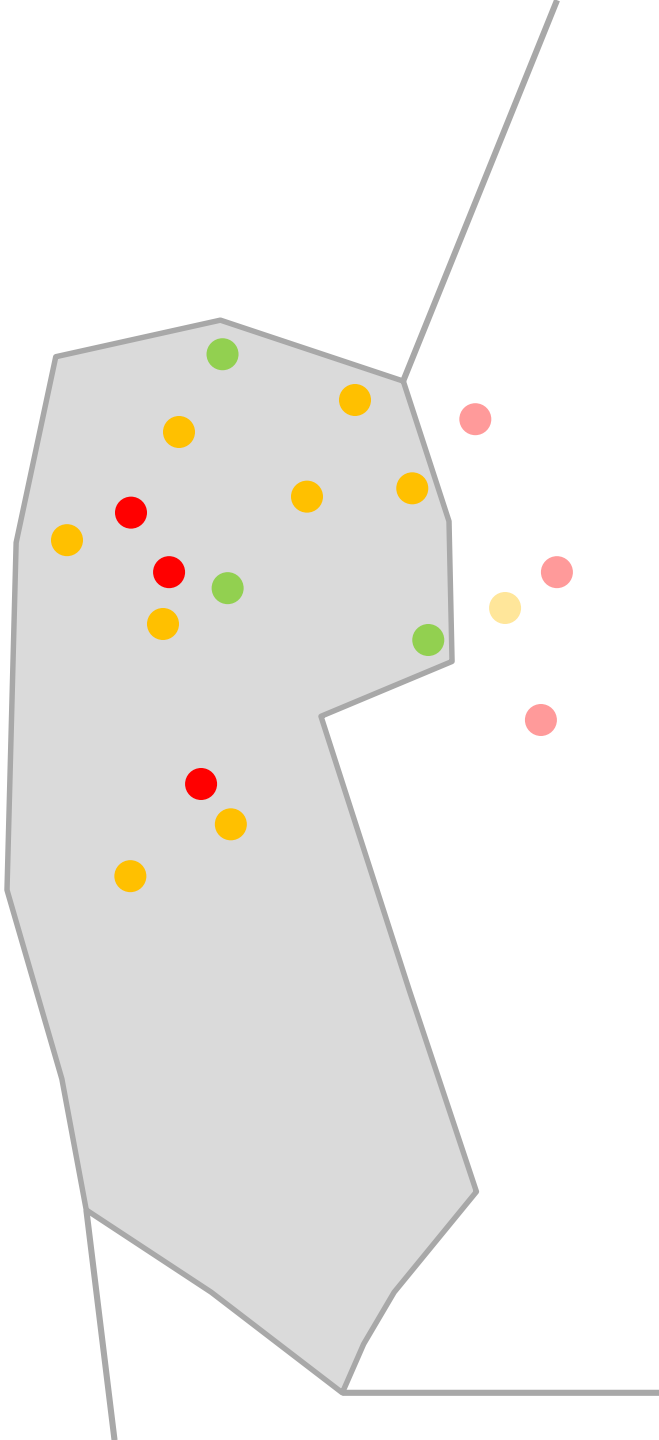

- Intensity of events is constant at all locations in the study area = homogenous
- Realizations from the model can appear different

# HPP vs IPP

- Assumption of stationary is restrictive– the intensity is almost never constant in the study area.
  - We expect more cases in areas with more people at risk
  - Risk may vary in space with other covariates (e.g geographic variations in population size / the amount of rainfall at a location)
- Heterogenous/inhomogenous Poisson process allows for spatially varying  first-order (mean) property of the process, where the intensity measure depends on other variables
- Non-stationary

# Edge effects

- Boundaries of edges of an area may be incompatible (arbitrary), unavailable (another country) or non-existent (next to an ocean)
- Points or area-units near edges will have fewer neighbors, which will be problematic when calculating mean intensity or test statistics for clustering.
- Corrections usually use a weighting system to give less weight to observations at the edge



# Global tests for spatial clustering using point (event) data

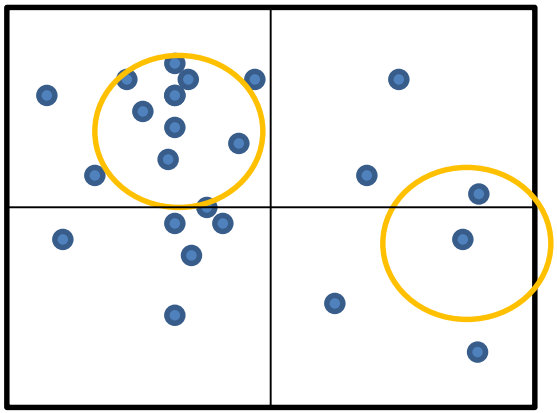
# Ripley's K function

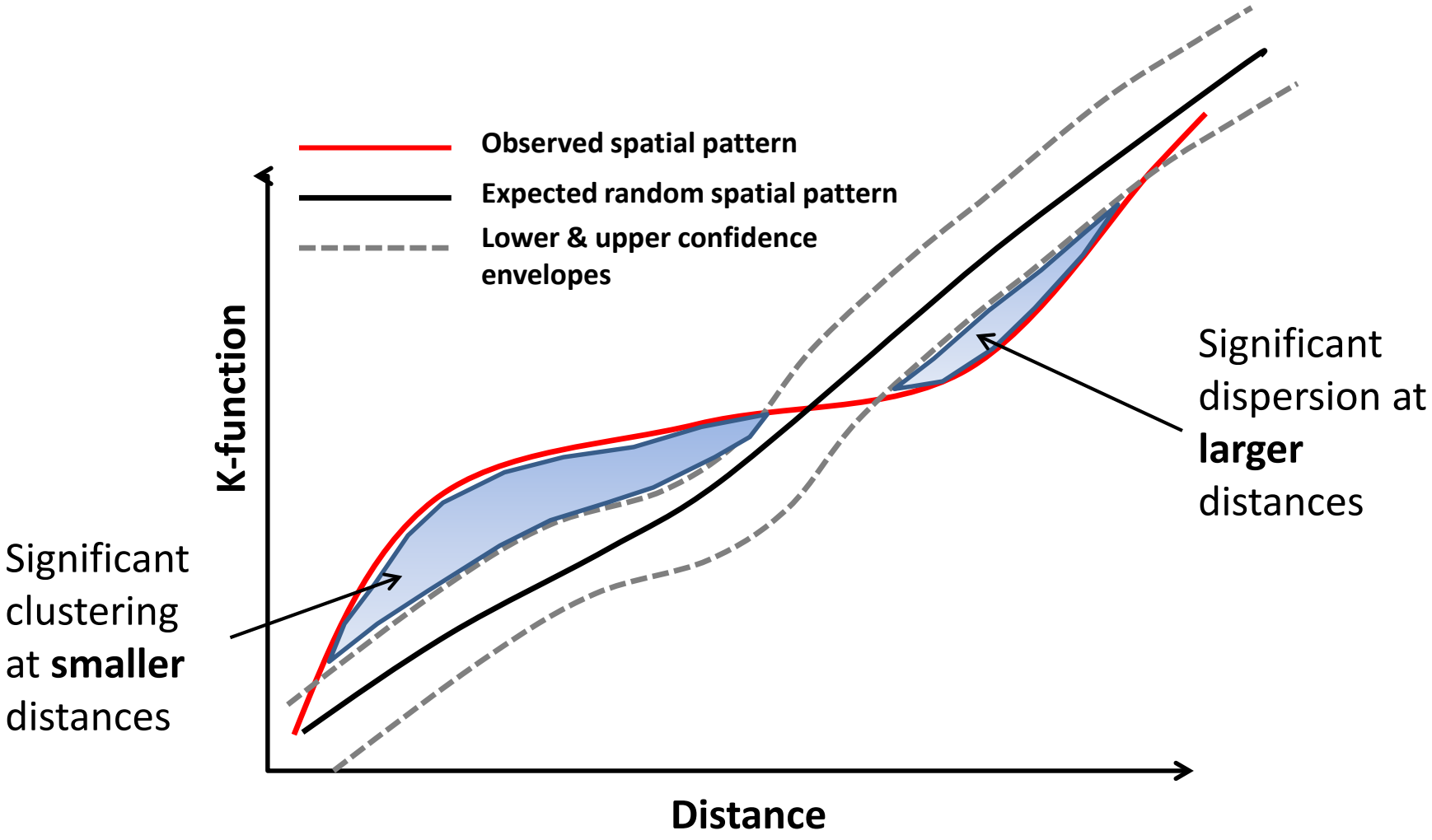
- Second-order analysis: Identifies the distance at which clustering occurs, based on the expected number of events within a distance, without regard to shape of study area and assuming no 1st order effects

$$K\ (h) = \frac{E[\text{number of events within } h \text{ of a } randomly \text{ chosen event}]}{\lambda}$$

- *lambda* is the average density (intensity) of points (expected number of points per unit area).

Λ = 23/4 = 5.75






When observed K value is larger than the expected K value for a particular distance, the distribution is more clustered than a random distribution at that distance. When the observed K value is smaller than the expected K, the distribution is more dispersed than a random distribution at that distance. Statistical significance indicated by envelopes

# Limitations

- Assumes stationary process
- K-function just for cases may not be useful due to variations in the spatial distribution of the population at risk

# Difference in K-functions

- Assess clustering of the cases compared to the controls
  - Calculate the K-function for both cases and non-cases (controls) and estimate the difference in the two functions to represent a measures of the extra aggregation of cases above that observed in the general population.
- $H_0$ = cases and controls are two IPP that have the same intensities up to a proportionality constant and so will produce the same K functions.

$$K_s = K_1(s) - K_0(s)$$

# Hypothesis testing

- Monte Carlo randomization can be used to randomly permute the locations of cases and non-cases (controls)
- Values of the difference function D(s) is computed for each permutation and then upper and lower bounds plotted.
- Can use ClusterSeer or the splancs package in R

# Example and in-class exercises

- Visualize data
- Use the case data characterize clustering using Ripley's K
  - Is clustering present in the dataset?
  - Over what scale is clustering present?
  - What are the limitations of this approach
- Estimate the difference in K-function for cases and controls

# Spatial scan statistics

- Calculations of intensity and K-functions not in many packages – other tests based on properties of heterogenous point process (i.e. number of events follows Poisson distribution)
- Scan statistics – based on comparisons of local rate estimates
  - Does the ratio of cases to controls (or cases per population at risk) appear “significantly” elevated in certain areas
  - Similar to the comparison of intensity functions
  - Goal is to detect specific clusters where the observed rate (or ratio) is inconsistent with the rate (or ratio) over the rest of the study area.

# Kulldorff's spatial scan statistic

This test detects spatial clusters based on a significant excess of cases within a moving cylindrical (or elliptical) window that systematically visits all spatial locations. Can be used for discrete (non-random and fixed locations) and continuous scan statistics (locations are random).

Different types of probabilistic models for discrete scan statistics according to the underlying distribution of the population at risk.

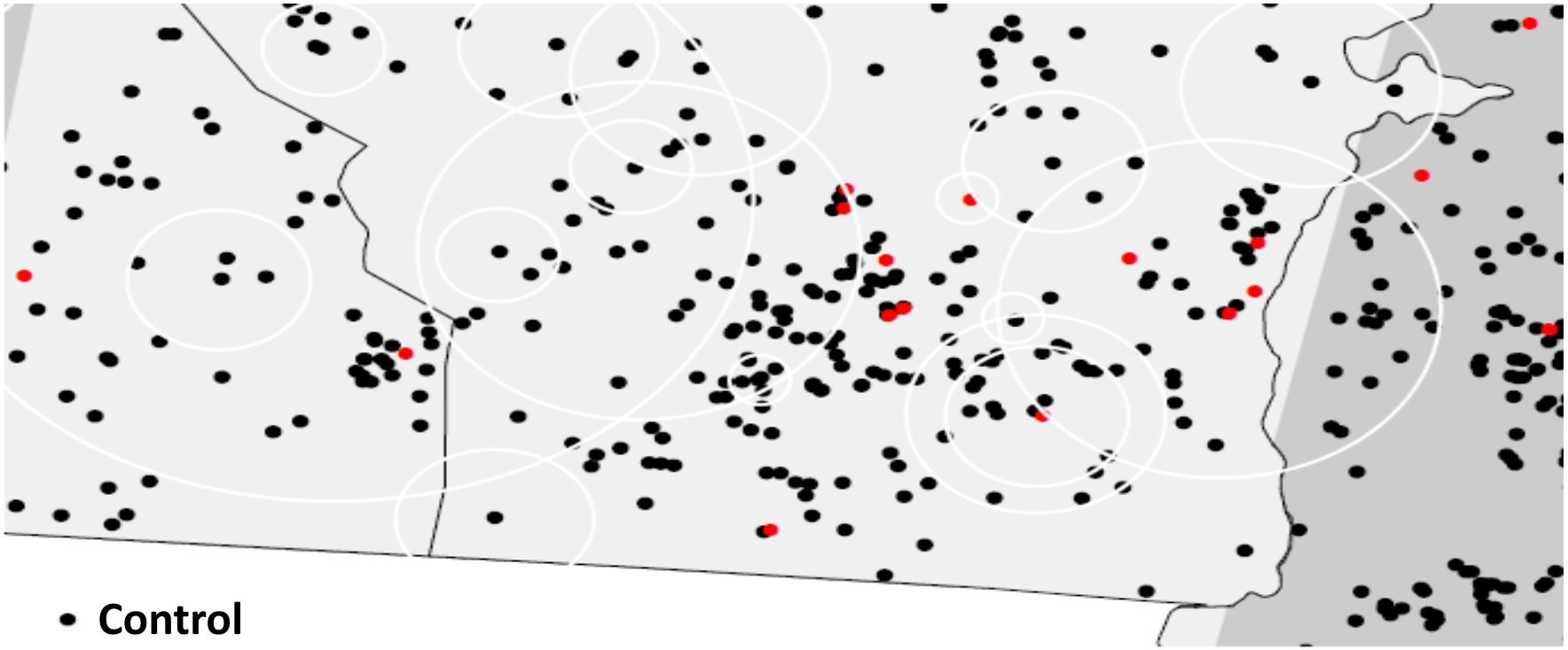
- Poisson: number of events in a geographical location is poisson-distributed, according to a known underlying population at risk
- Bernoulli: 0/1 event data such as cases and controls
- Ordinal: Ordered categorical data
- Exponential: Survival time data
- Normal: Other types of continuous data

**$H_1$ :** ***there is an elevated risk within the window as compared to outside, for each location and size.***

**$H_0$:** ***The risk of an event is the same within and outside the window***

- Create a grid of points (either as grid or population unit locations) throughout area

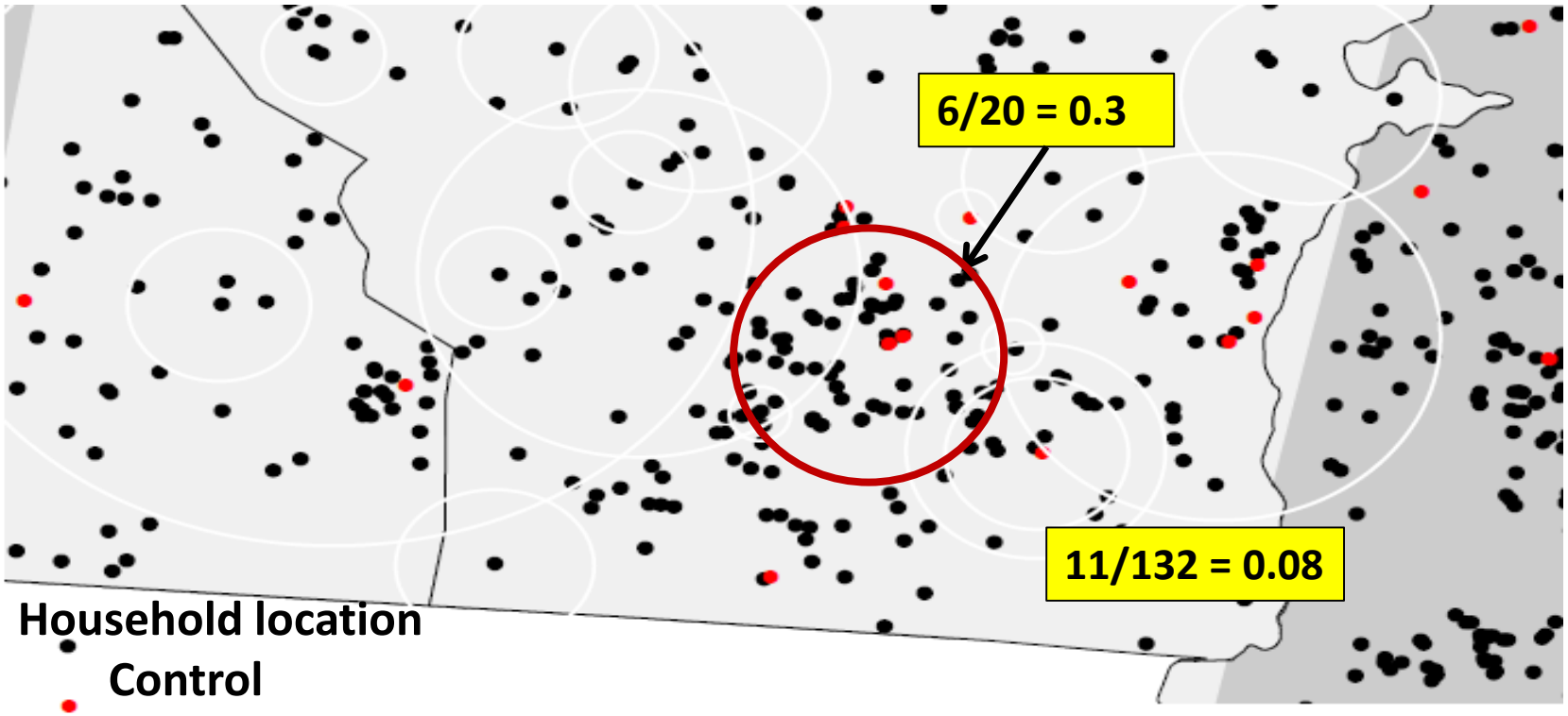
- Creates an infinite number of circles of varying sizes around each point, maximum size = fixed proportion of pop.



- **Control**
- **Malaria case**

**For each circle:**

- Obtain actual and expected number of events inside and outside the circle
- Expected number of cases in each unit is proportional to the size of the population at risk
- Calculate the likelihood function – likelihood of finding the observed number of events in each circle'



# Data input

- Data can be either aggregated by administrative unit or other geographical level, or there may be unique coordinates for each observation.
- SaTScan™ adjusts for the underlying spatial variation of a background population.
- It can also adjust for any number of categorical covariates, as well as for temporal trends, known space-time clusters and missing data.

# Common SatScan models

- **Poisson model:** where the number of events in an area is Poisson distributed under the null hypothesis. Number of cases is compared to the background population data and the expected number of cases is proportional to the size of the population at risk in each circle.
- **Bernoulli model**: with 0/1 event data such as cases and controls, where non-cases are taken to represent the background distribution population.

# In-class assignment: SatScan analysis

- Do the SatScan practical using the case control data
- Compare SatScan results using spscan.test in the “smacpod” library in R

NB: there is now an rsatscan package that runs SatScan from R.