Lesson 4: Spatial Point Patterns Analysis

Dr. Kam Tin Seong
Assoc. Professor of Information Systems(Practice)
School of Computing and Information Systems,
Singapore Management University

29 Jan 2023

Content

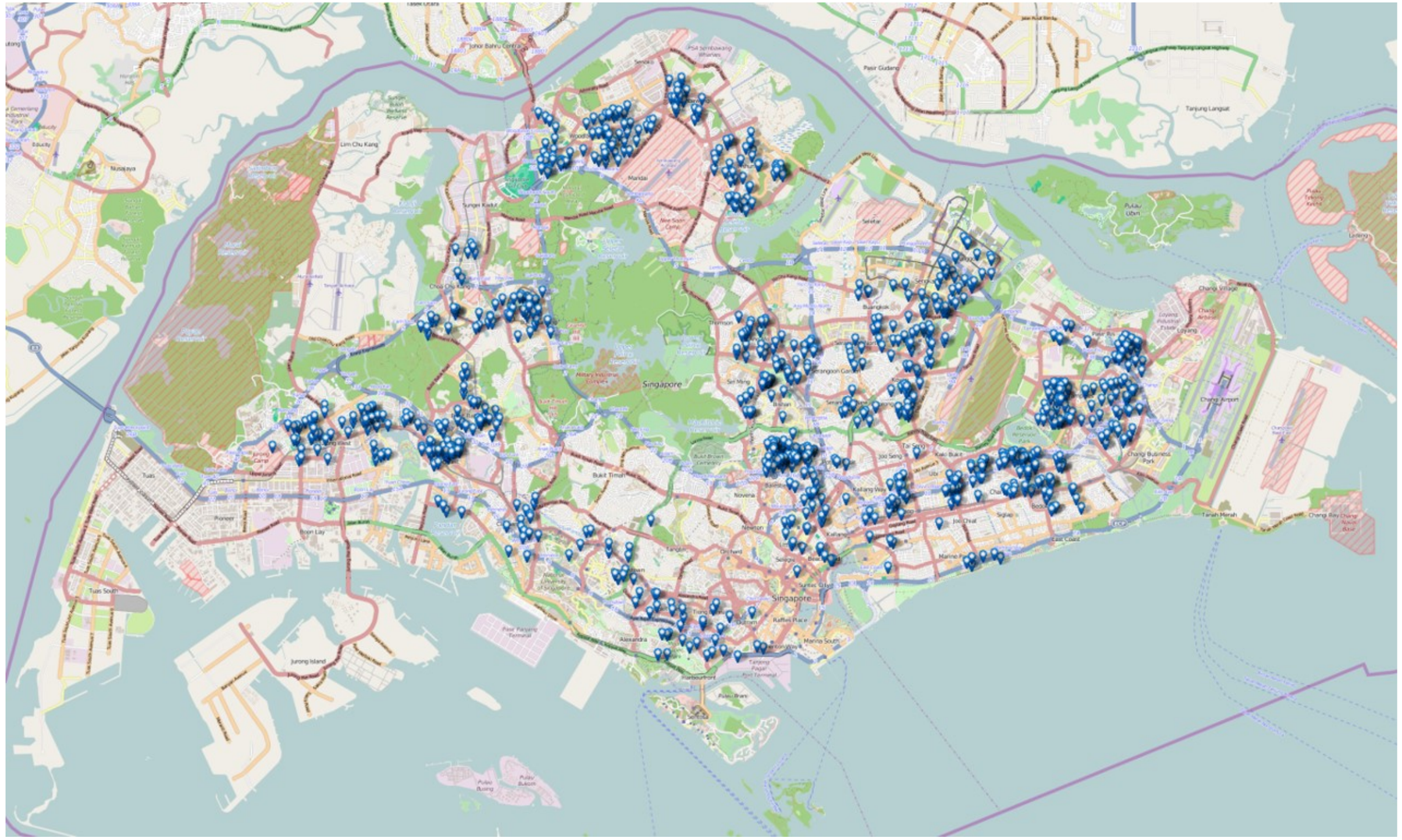
- Introducing Spatial Point Patterns
  - The basic concepts of spatial point patterns
  - 1st Order versus 2nd Order
  - Spatial Point Patterns in real world
- 1st Order Spatial Point Patterns Analysis
  - Quadrat analysis
  - Kernel density estimation
- 2nd Order Spatial Point Patterns Analysis
  - Nearest Neighbour Index
  - G-function
  - F-function
  - K-function
  - L-function

What is Spatial Point Patterns

- Points as Events
- Mapped pattern
  - Not a sample
  - Selection bias
- Events are mapped, but non-events are not

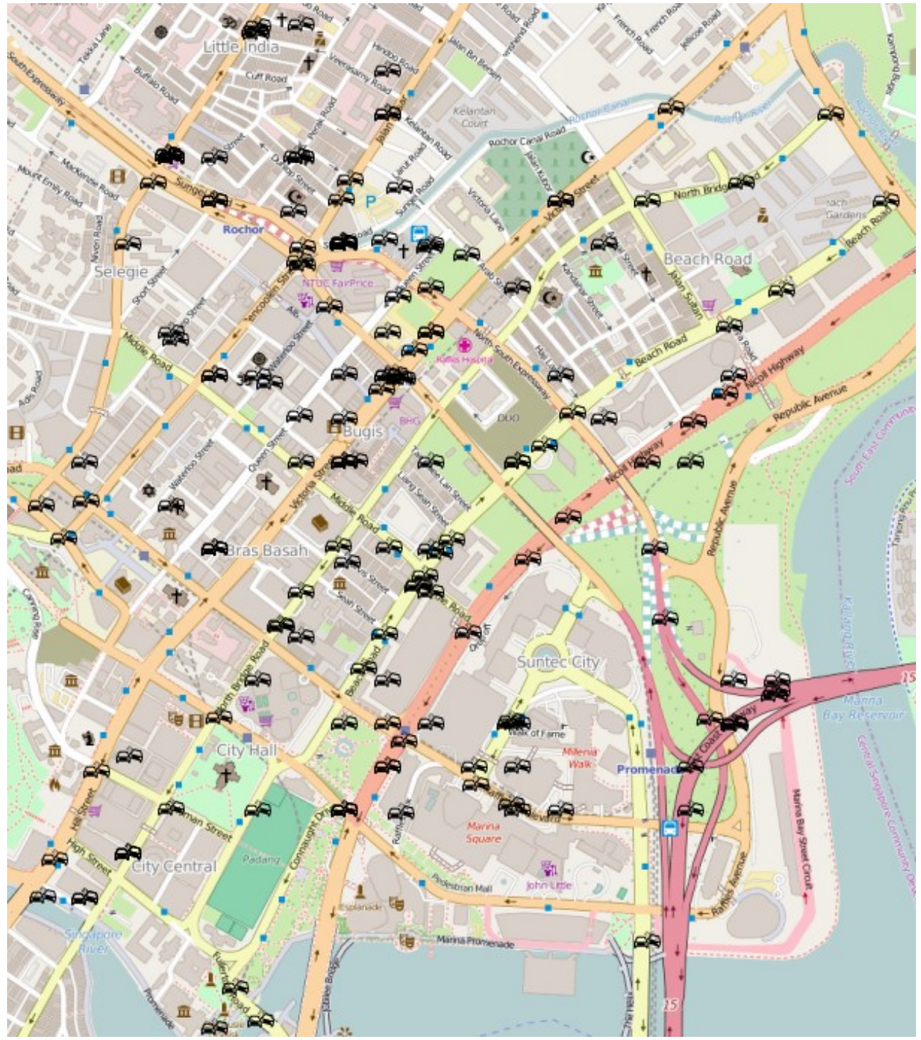
Spatial Point Patterns in Real World

- Distribution of dieses such as dengue fever.



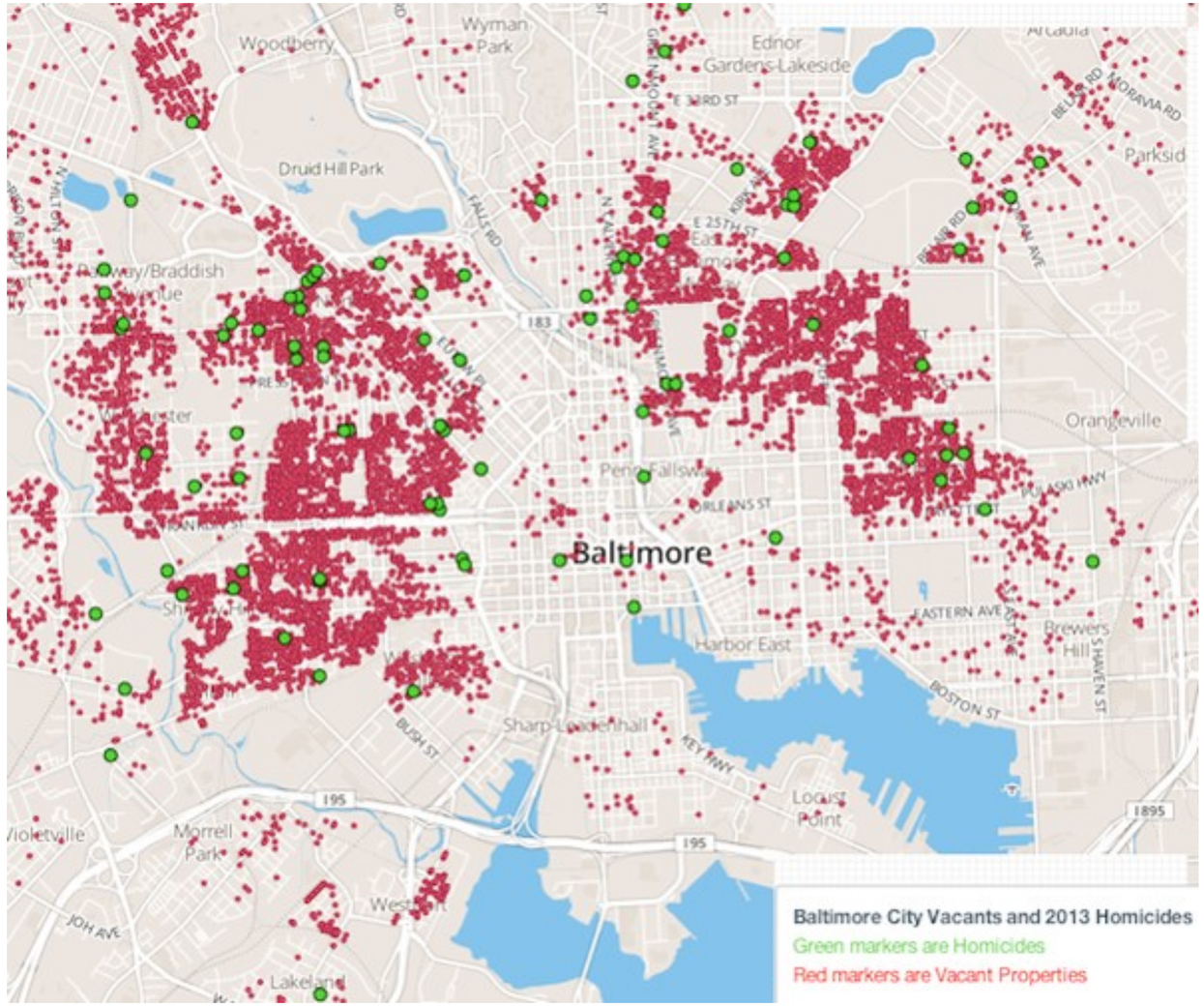
Spatial Point Patterns in Real World

- Distribution of car collisions.



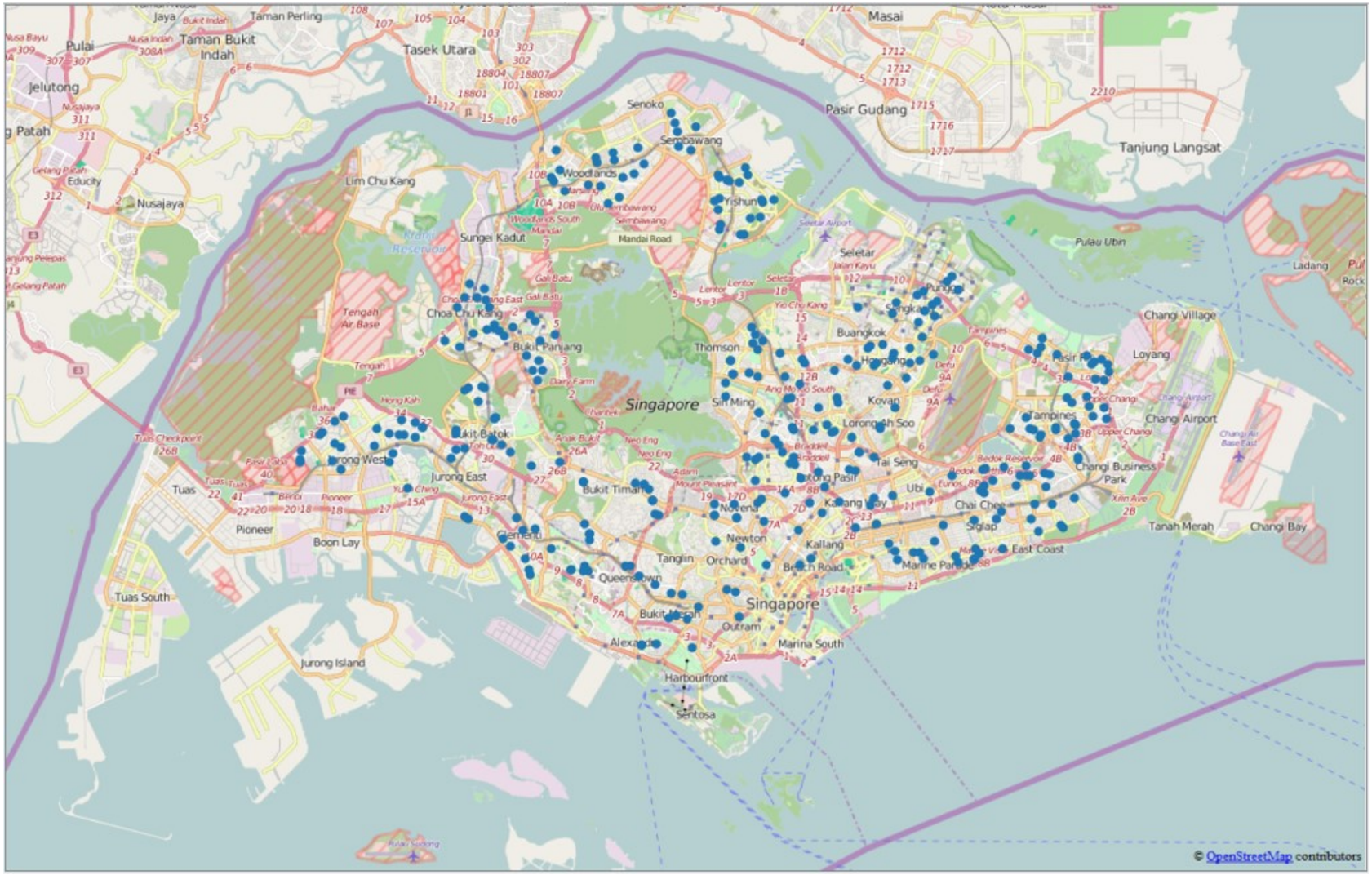
Spatial Point Patterns in Real World

- Distribution of crime incidents.



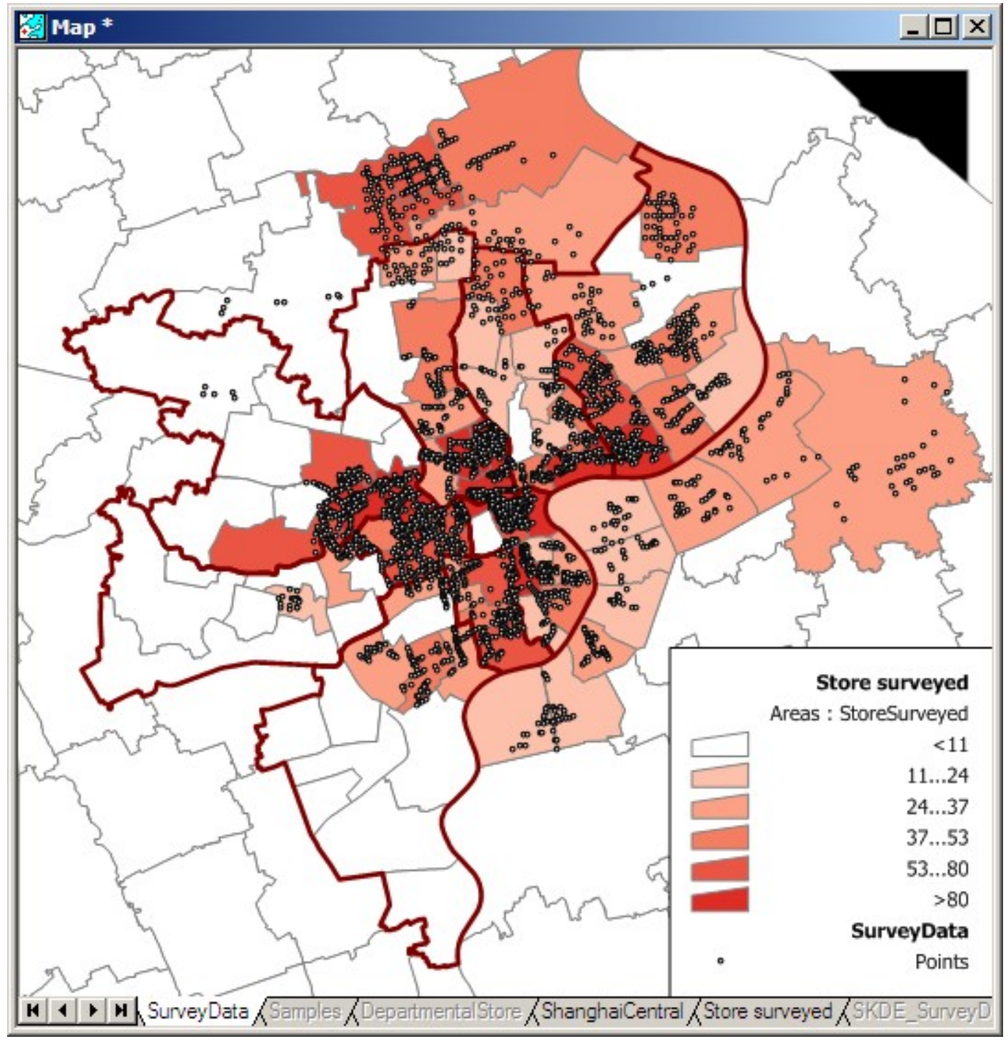
Spatial Point Patterns in Real World

- Distribution of public services such as education institutions.



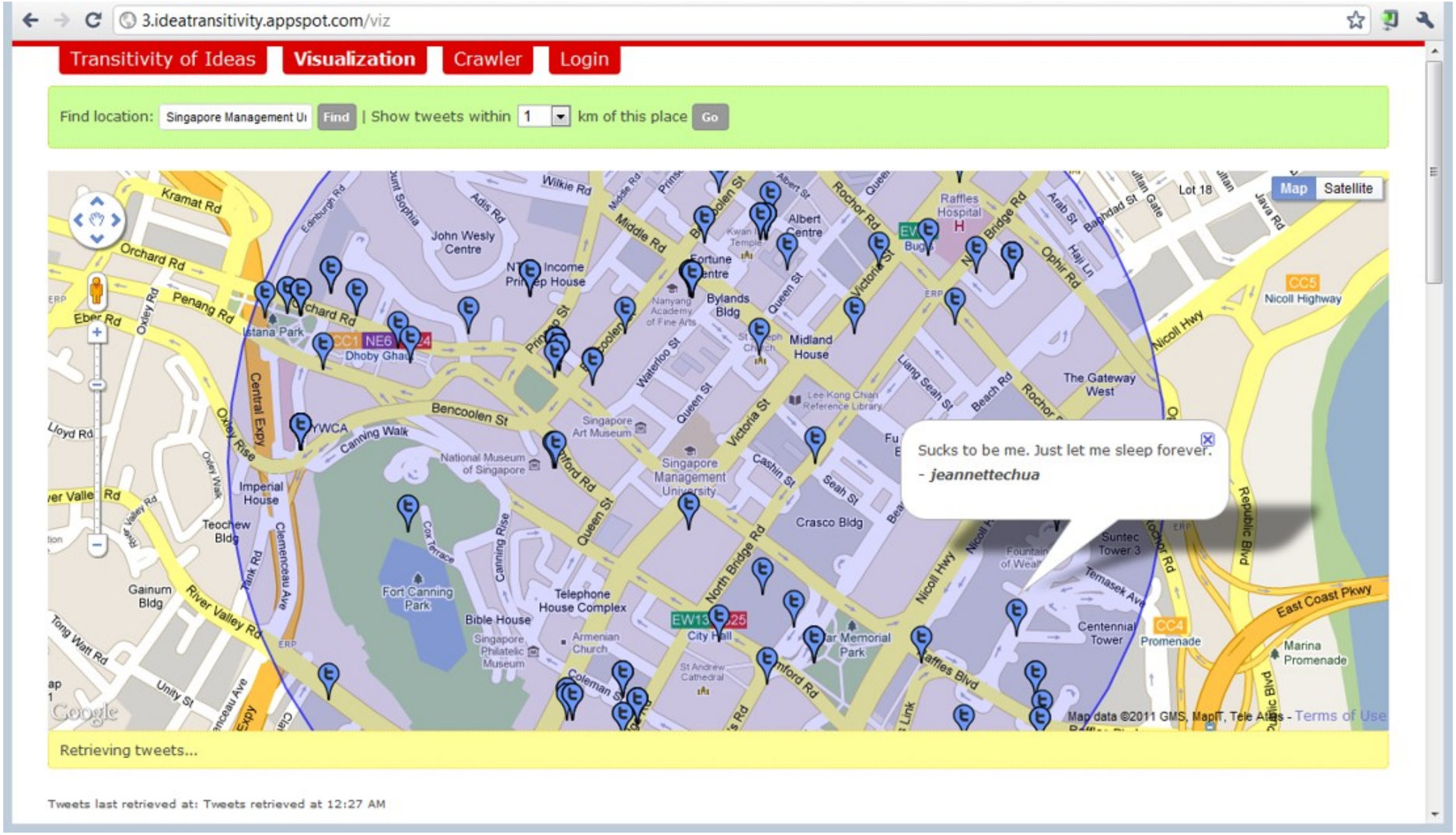
Spatial Point Patterns in Real World

- Locations of the different channel stores.



Spatial Point Patterns in Real World

- Distribution of social media data such as tweets.



Real World Question

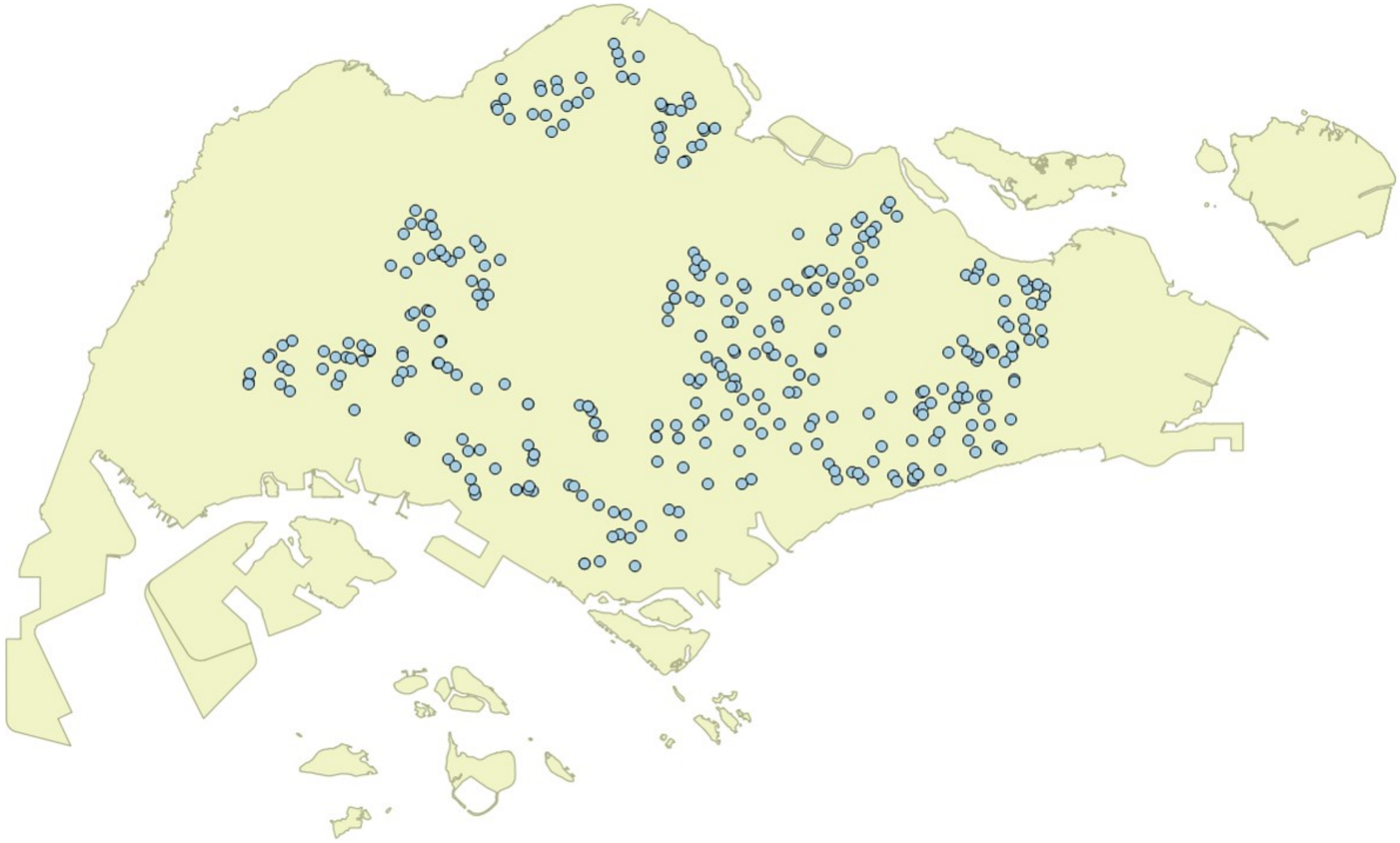
- Location only
  - are points randomly located or patterned
- Location and value
  - marked point pattern
  - is combination of location and value random or patterned
- What is the underlying process?

Points on a Plane

- Classic point pattern analysis
  - points on an isotropic plane
  - no effect of translation and rotation
  - classic examples: tree seedlings, rocks, etc
- Distance
  - straight line only

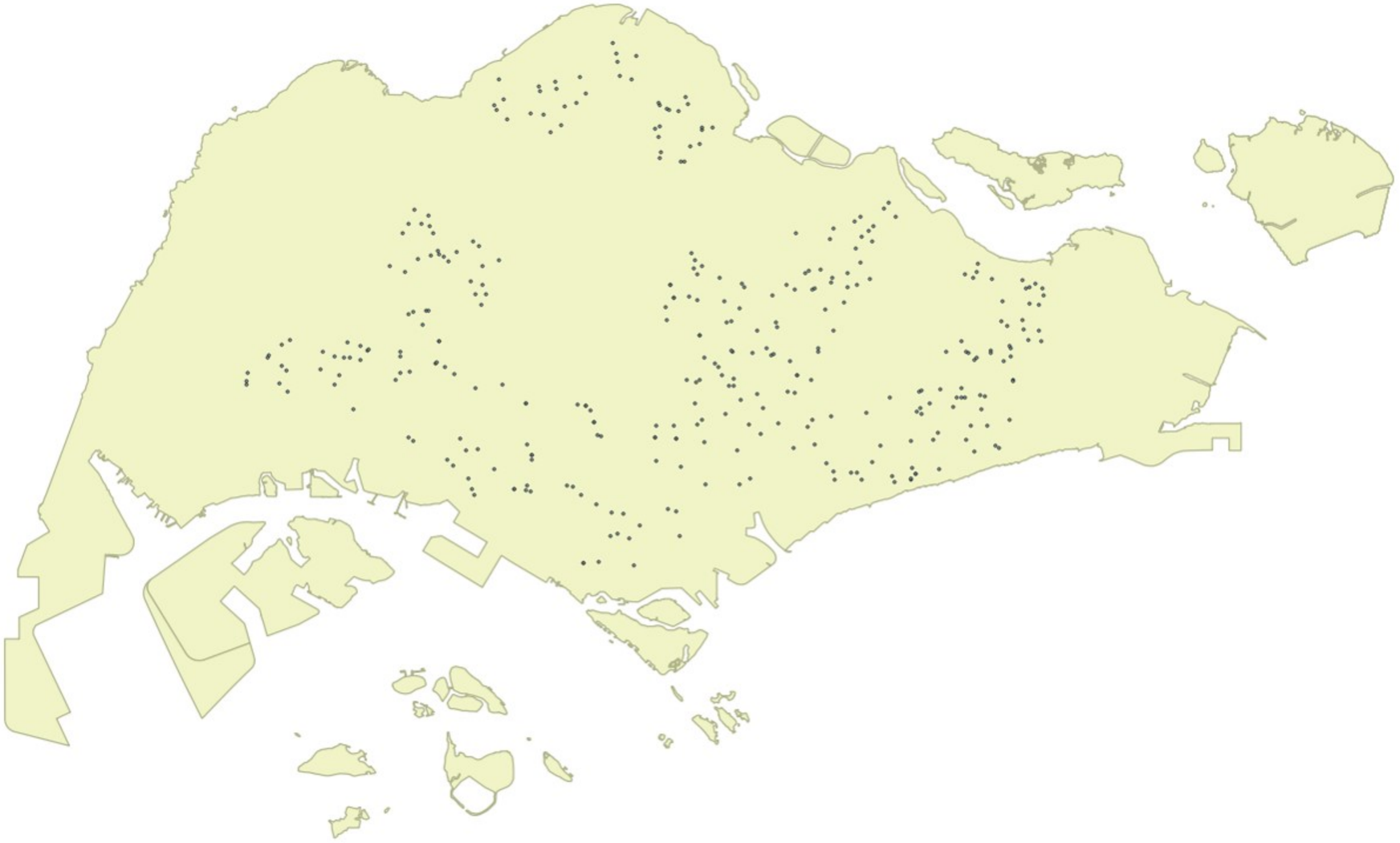
Real world spatial point patterns

- Is this a random distribution?



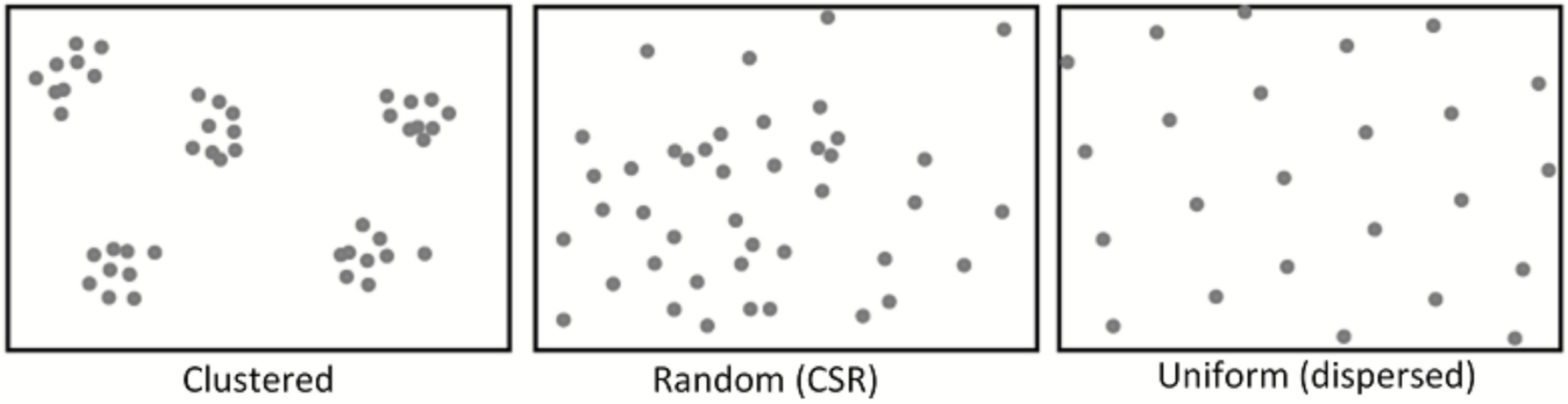
Real world spatial point patterns

- Is this a random distribution?



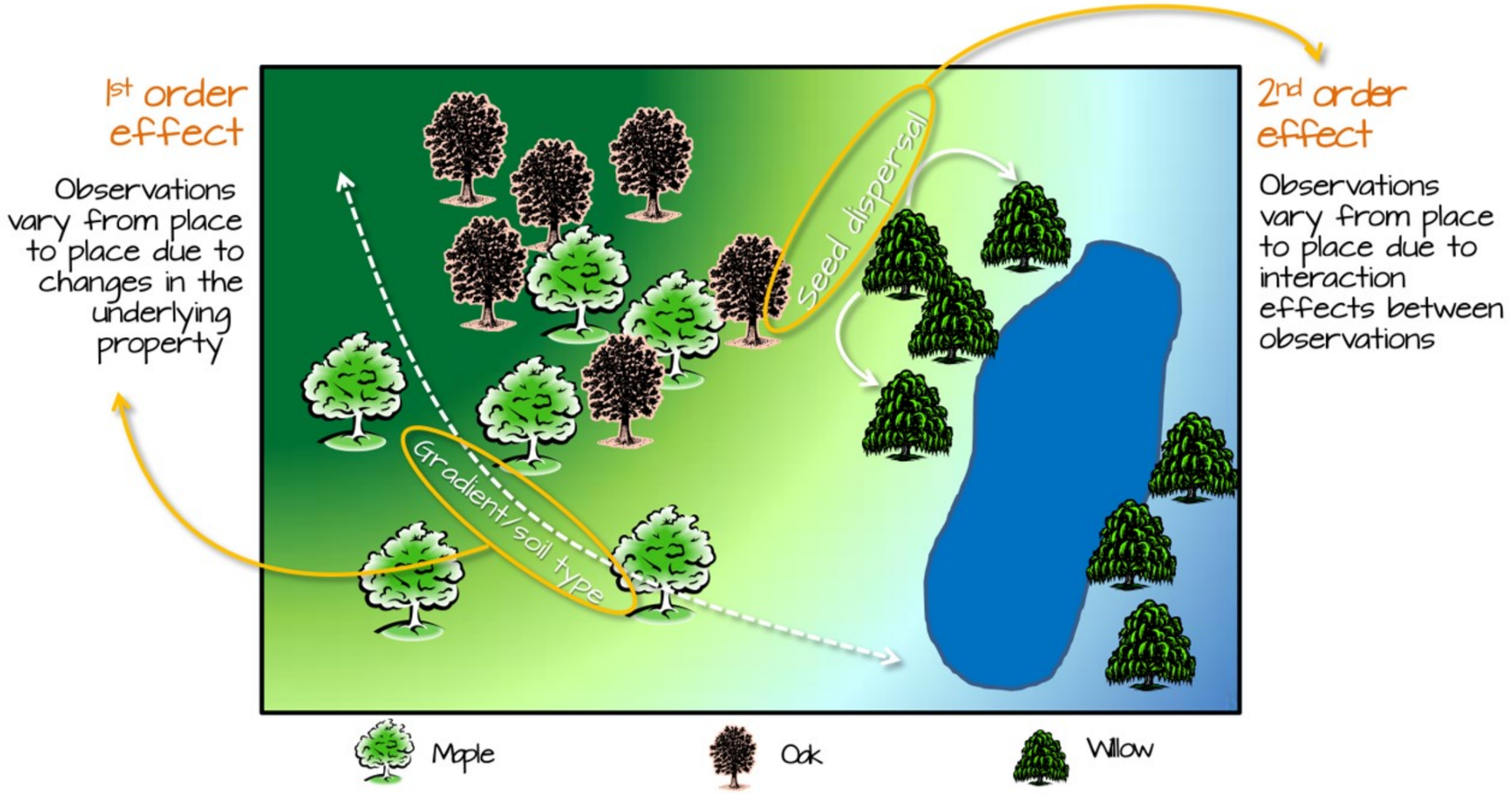
Spatial Point Patterns Analysis

- Point pattern analysis (PPA) is the study of the spatial arrangements of points in (usually 2-dimensional) space.
- The simplest formulation is a set $X = \{x \in D\}$ where D, which can be called the **study region**, is a subset of Rn, a n-dimensional **Euclidean space**.
- A fundamental problem of PPA is inferring whether a given arrangement is merely **random** or the result of some process.



Spatial Point Patterns Analysis Techniques

- First-order vs Second-order Analysis of spatial point patterns.



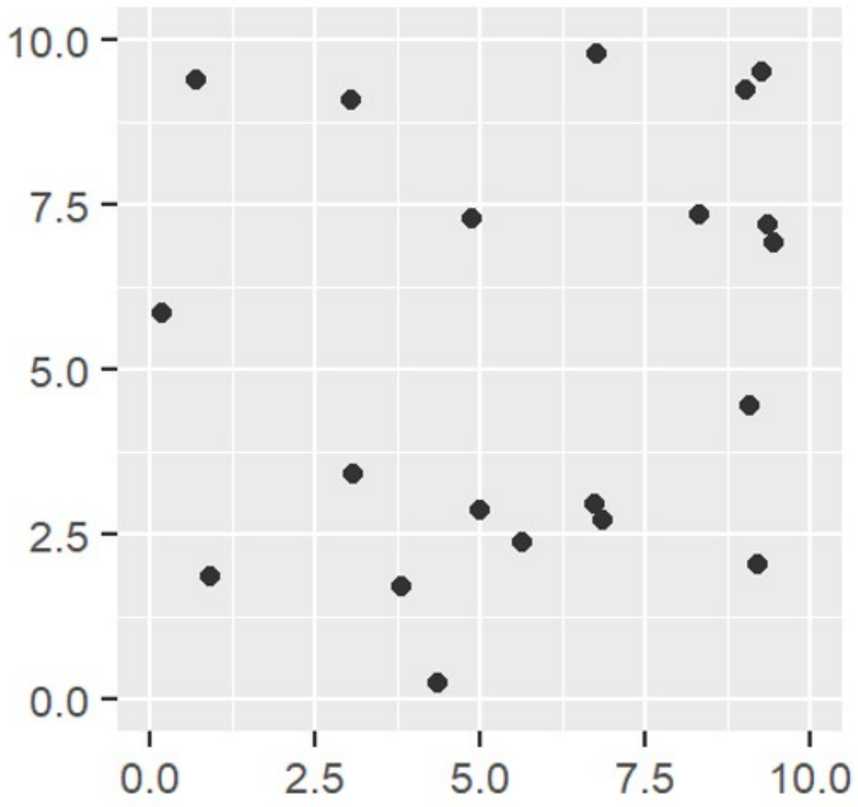
Reference: 11.4 First and second order effects of Intro to GIS and Spatial Analysis

First-order Spatial Point Patterns Analysis Techniques

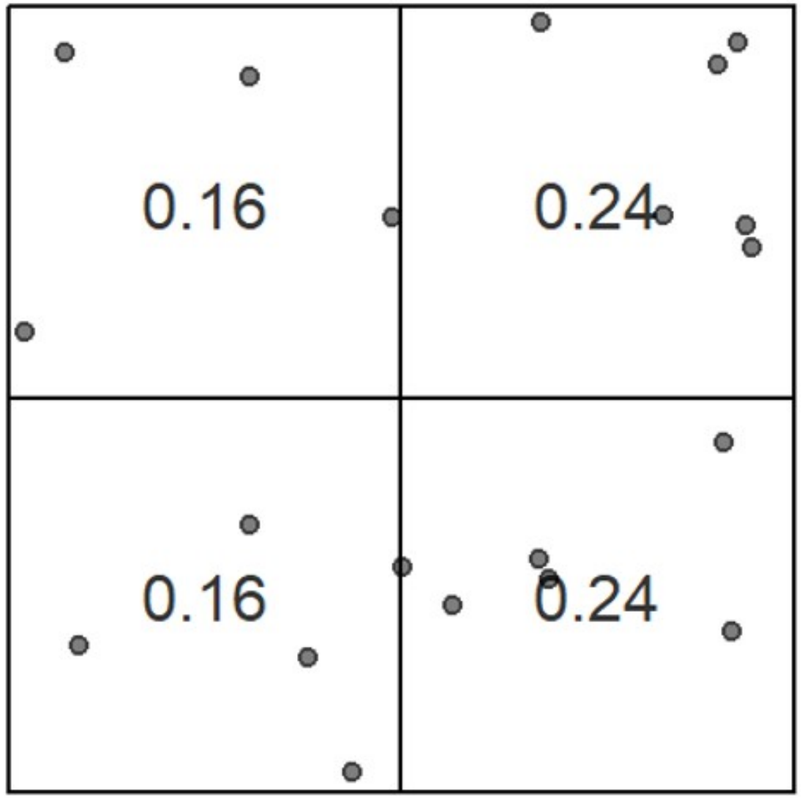
- Density-based
  - Kernel density estimation
  - Quadrat analysis,
- Distance-based
  - Nearest Neighbour Index

Basic concept of density-based measures

Global density



Local density



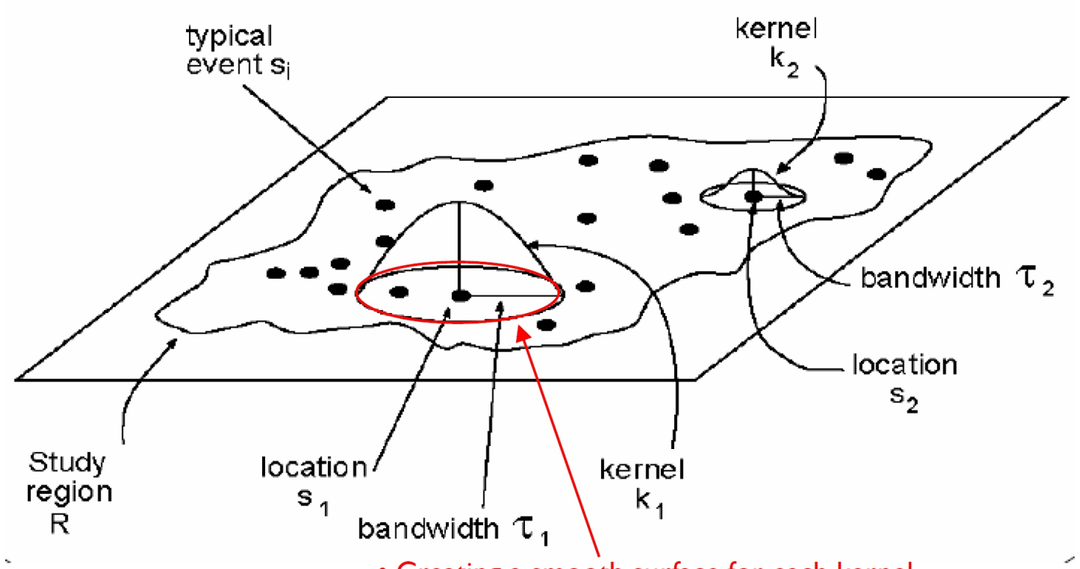
Kernel density estimation (Silverman 1986)

- A method to compute the intensity of a point distribution.

The general formula:

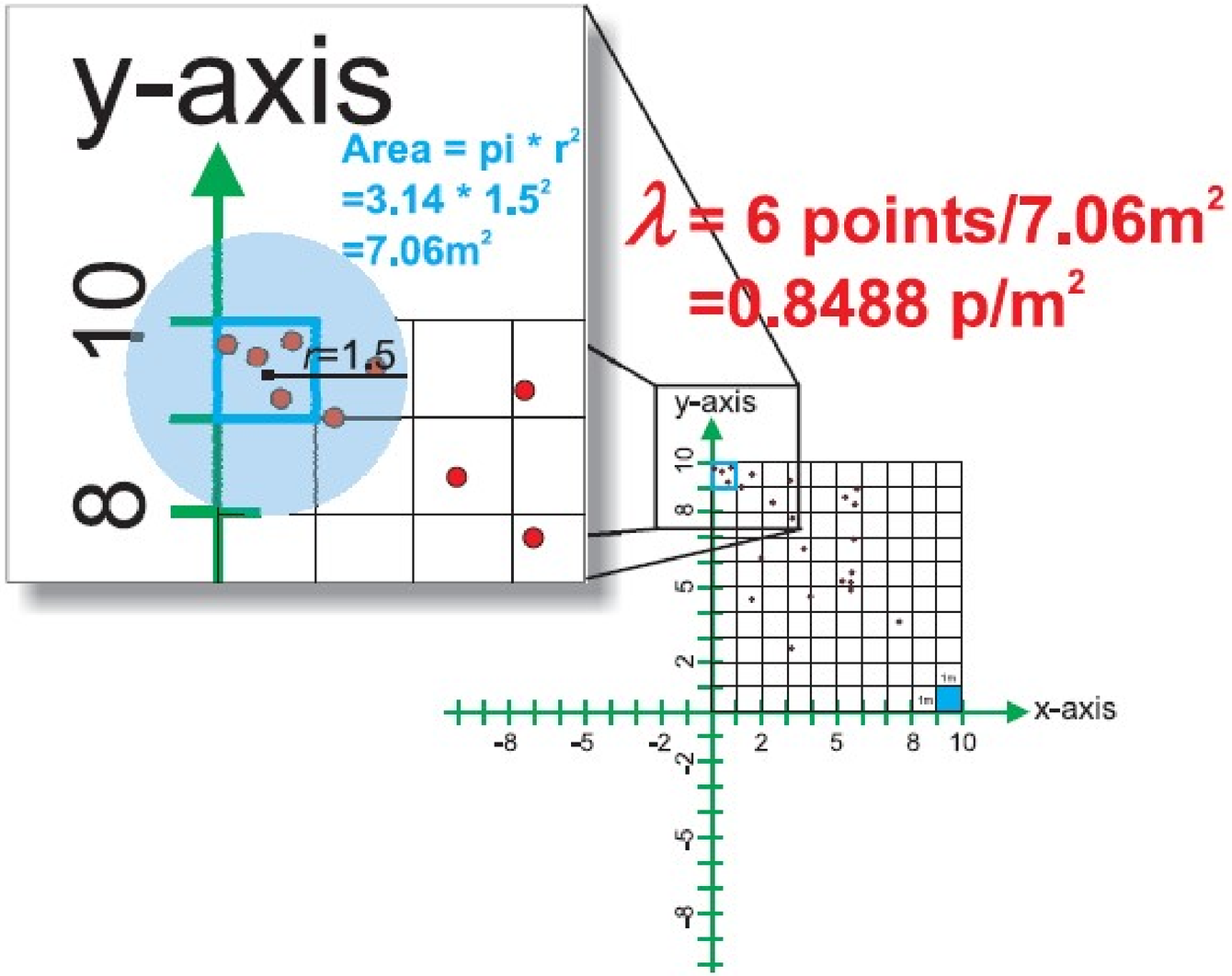
$$\hat{\lambda}_\tau(s) = \frac{1}{\sigma_\tau(s)} \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right)$$

Graphically



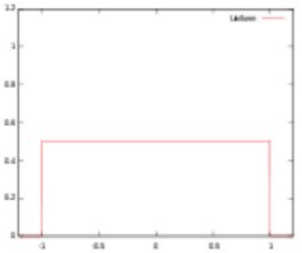
- Creating a smooth surface for each kernel
- Surface value highest in the center (point location) and diminishes with distance…reaches 0 at radius distance

KDE Step 1: Computing point intensity

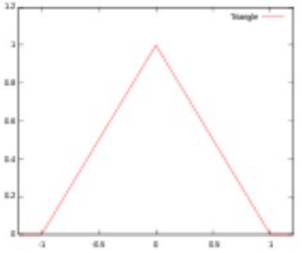


KDE Step 2: Spatial interpolation using kernel function

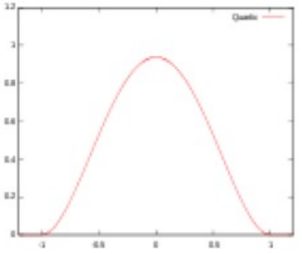
Uniform



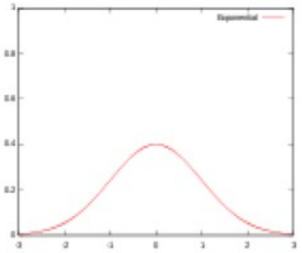
Triangular



Quartic



Gaussian



$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right)$$

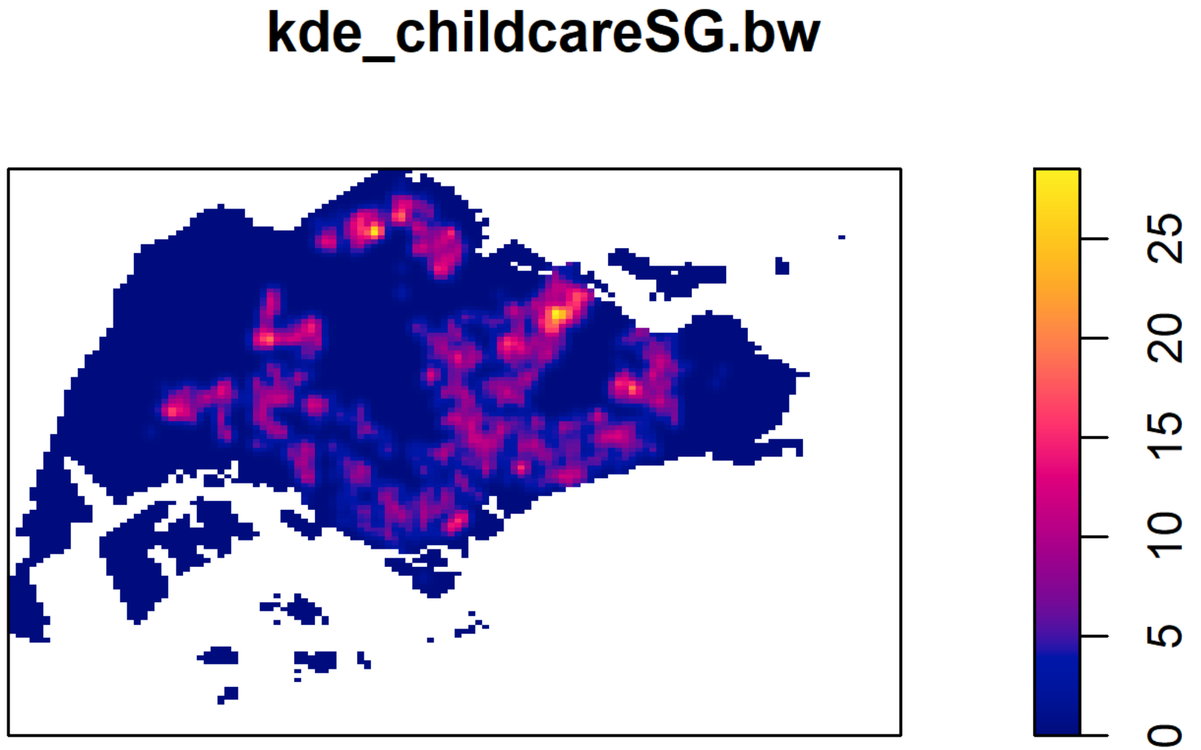
Each kernel type has a different equation for the function k, for example:

Triangular: $k = 1 - \left|\frac{d_i}{\tau}\right|$

Quartic: $k = \frac{3}{\pi}\left(1 - \frac{h_i^2}{\tau^2}\right)$

Normal: $k = \frac{1}{\sqrt{2\pi}} e^{-\frac{h_i^2}{2\tau^2}}$

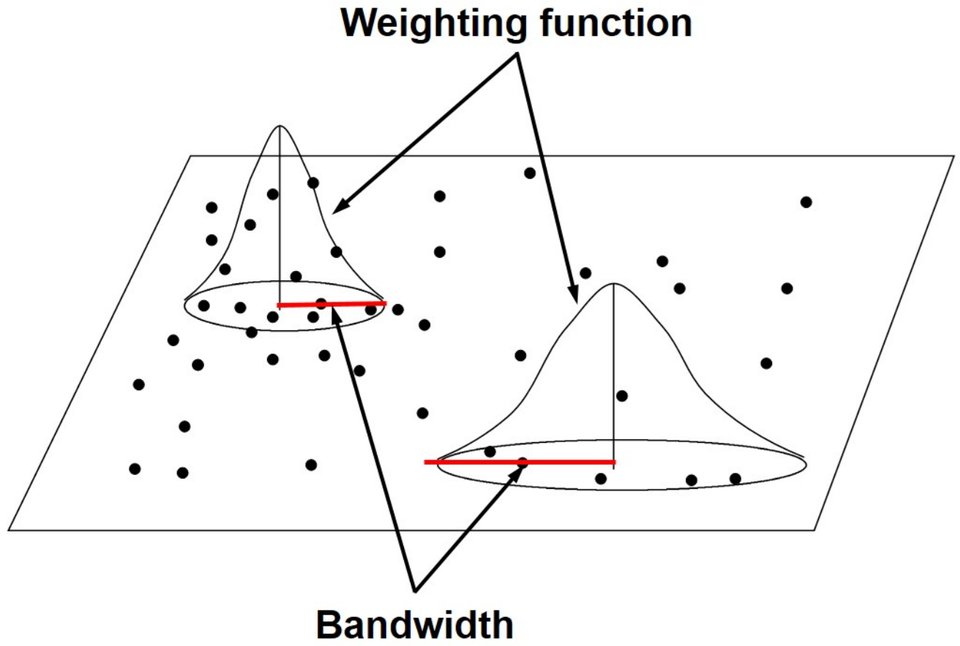
KDE Map of Childcare Services, Singapore



Adaptive Bandwidth

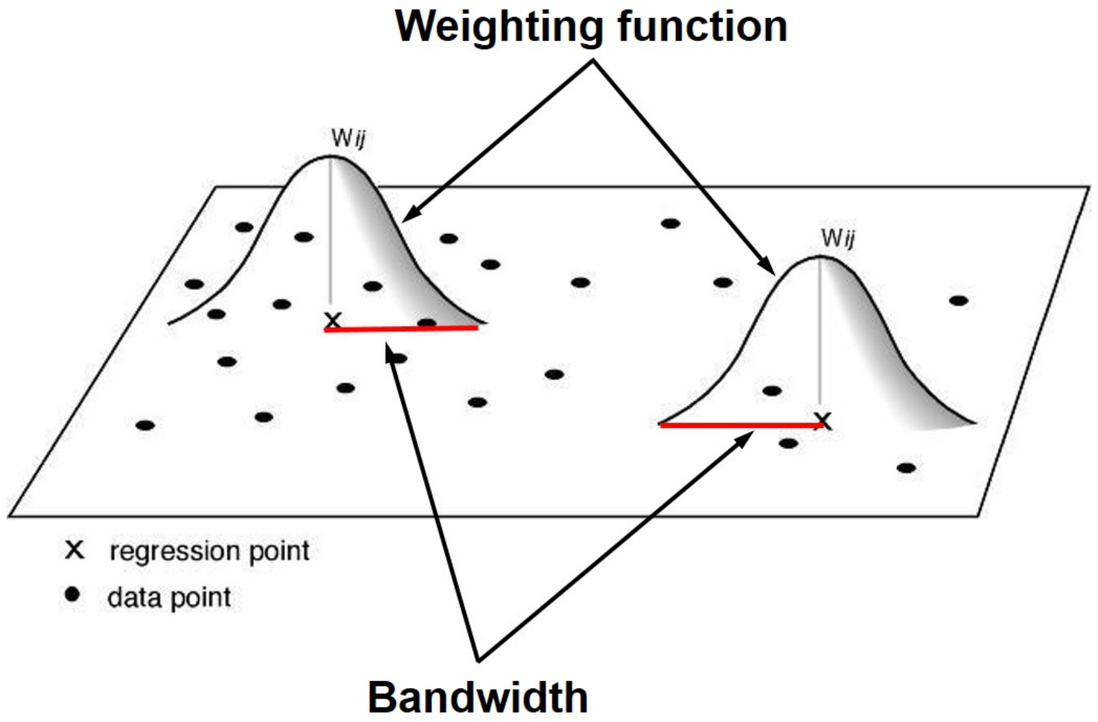
Adaptive schemes adjust itself according to the density of data: - Shorter bandwidths where data are dense and longer where sparse.

- Finding nearest neighbors are one of the often used approaches.



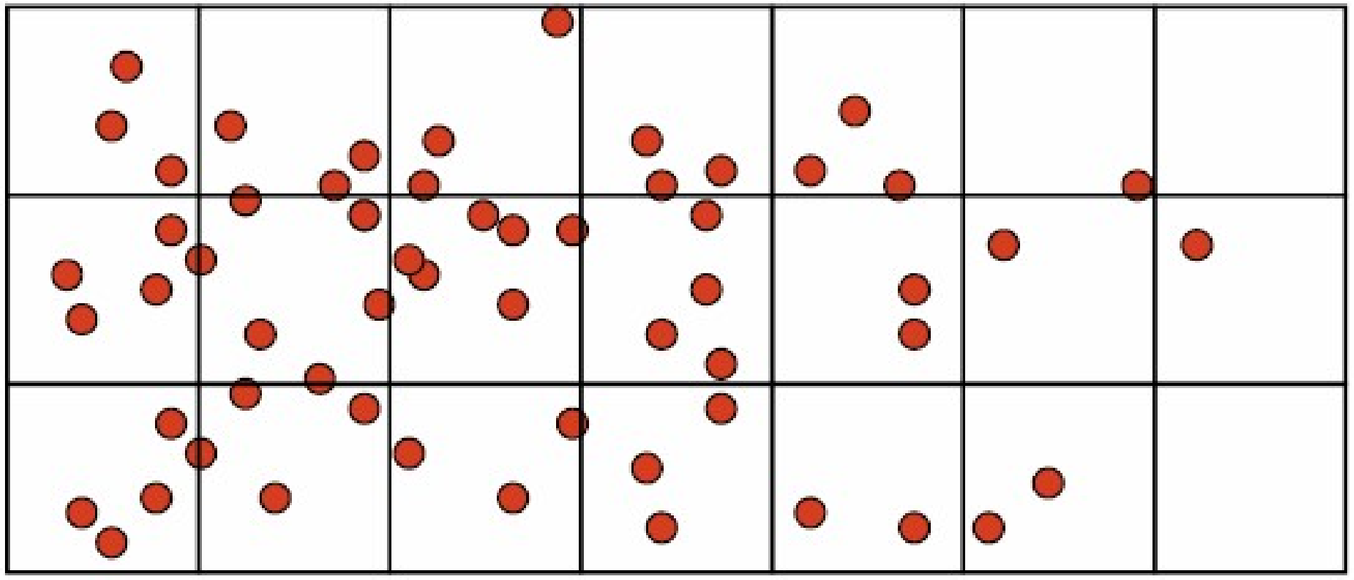
Fixed bandwidth

- Might produce large estimate variances where data are sparse, while mask subtle local variations where data are dense.
- In extreme condition, fixed schemes might not be able to calibrate in local areas where data are too sparse to satisfy the calibration requirements (observations must be more than parameters).



Quadrat Analysis – Step 1

- Divide the study area into subregion of equal size,
  - often squares, but don’t have to be.



Quadrat Analysis – Step 2

- Count the frequency of events in each region.

	3	3	3	3	3	1	0
number	4	6	6	4	2	1	1
	4	4	3	3	2	2	0

Quadrat Analysis – Step 3

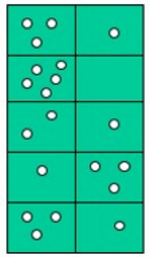
- Calculate the intensity of events in each region.

	.75	.75	.75	.75	.75	.25	0
Intensity λ	1	1.5	1.5	1	.5	.25	.25
	1	1	.75	.75	.5	.5	0

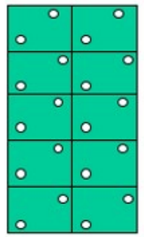
$\lambda = n/A$ where n =number of events and A = 4 is area of each quadrat

Quadrat Analysis – Step 4

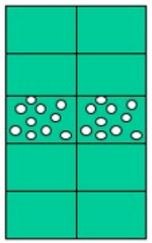
- Calculate the quadrat statistics and perform CSR test.



RANDOM



UNIFORM/
DISPERSED



CLUSTERED

3	1
5	0
2	1
1	3
3	1

Quadrat #	# of Points Per Quadrat	x^2
1	3	9
2	1	1
3	5	25
4	0	0
5	2	4
6	1	1
7	1	1
8	3	9
9	3	9
10	1	1
	20	60
Variance	2.222	
Mean	2.000	
Var/Mean	1.111	

2	2
2	2
2	2
2	2
2	2

Quadrat #	# of Points Per Quadrat	x^2
1	2	4
2	2	4
3	2	4
4	2	4
5	2	4
6	2	4
7	2	4
8	2	4
9	2	4
10	2	4
	20	40
Variance	0.000	
Mean	2.000	
Var/Mean	0.000	

0	0
0	0
10	10
0	0
0	0

Quadrat #	# of Points Per Quadrat	x^2
1	0	0
2	0	0
3	0	0
4	0	0
5	10	100
6	10	100
7	0	0
8	0	0
9	0	0
10	0	0
	20	200
Variance	17.778	
Mean	2.000	
Var/Mean	8.889	

$$\text{N} = \text{number of quadrats} = 10$$

$$Variance = \frac{\sum x^2 - [(\sum x)^2 / N]}{N-1}$$

$$\text{Variance mean ratio} = \frac{\text{variance}}{\text{mean}}$$

To test for CSR, calculate the test statistic for quadrat (χ^2):

$$= \frac{(m-1)s^2}{\bar{x}}$$

m = # of quadrats
s^2 = observed variance
x = observed mean

Compare to χ^2 distribution with m-1 degrees of freedom

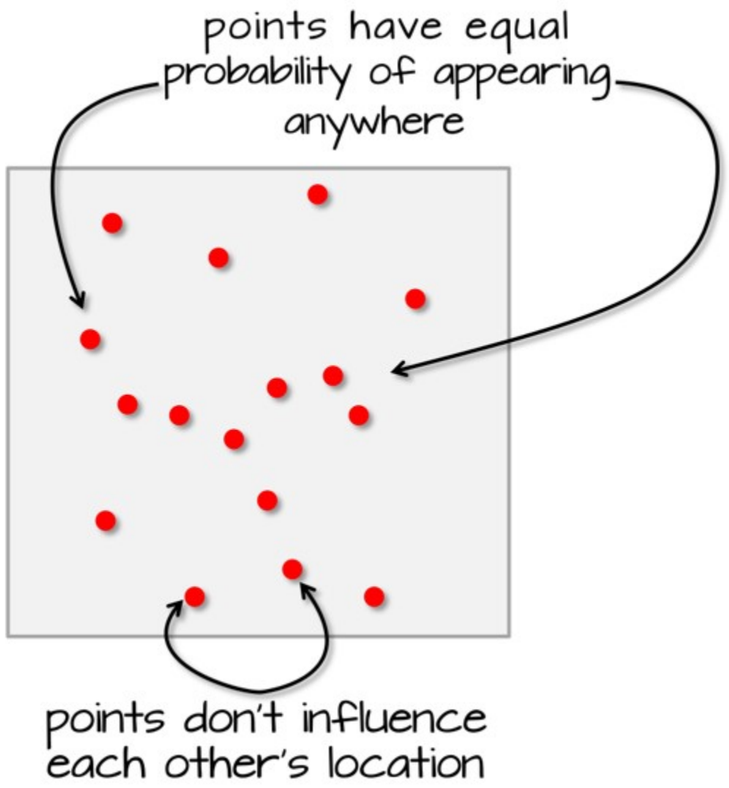
Quadrat Analysis – Variance-Mean Ratio (VMR)

- For an **uniform** distribution, the variance is zero, - therefore, we expect a variance-mean ratio **close to 0**.
- For a **random** distribution, the variance and mean are the same,
  - therefore, we expect a variance-mean ratio **close to 1**.
- For a **cluster** distribution, the variance is relatively large,
  - therefore, we expect a variance-mean ratio **greater than 1**.]

Complete Spatial Randomness (CSR)

- CSR/IRP satisfy two conditions: - Any event has equal probability of being in any location, a **1st order** effect.
  - The location of one event is independent of the location of another event, a **2nd order** effect.

Reference: Chapter 12 Hypothesis testing of Intro to GIS and Spatial Analysis



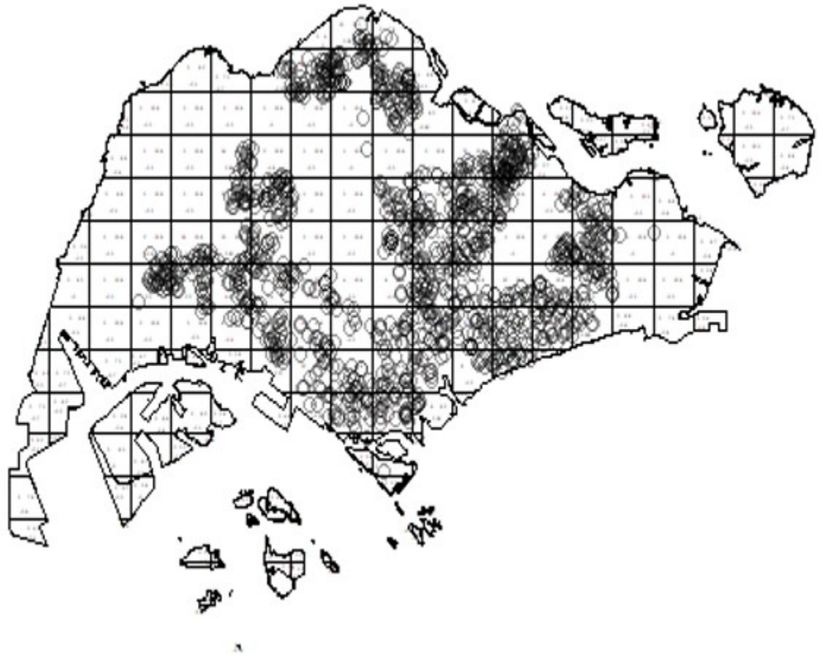
Quadrat Analysis: The interpretation

```
        Chi-squared test of CSR using quadrat counts
        Pearson X2 statistic

data:  sp_ppp
X2 = 245.23, df = 23, p-value < 2.2e-16
alternative hypothesis: two.sided

Quadrats: 6 by 4 grid of tiles
```

The Chi-squared statistic is large and the p-value is smaller than 0.05 => Reject the null hypothesis that the point patterns are randomly distributed.

Weaknesses of quadrat analysis

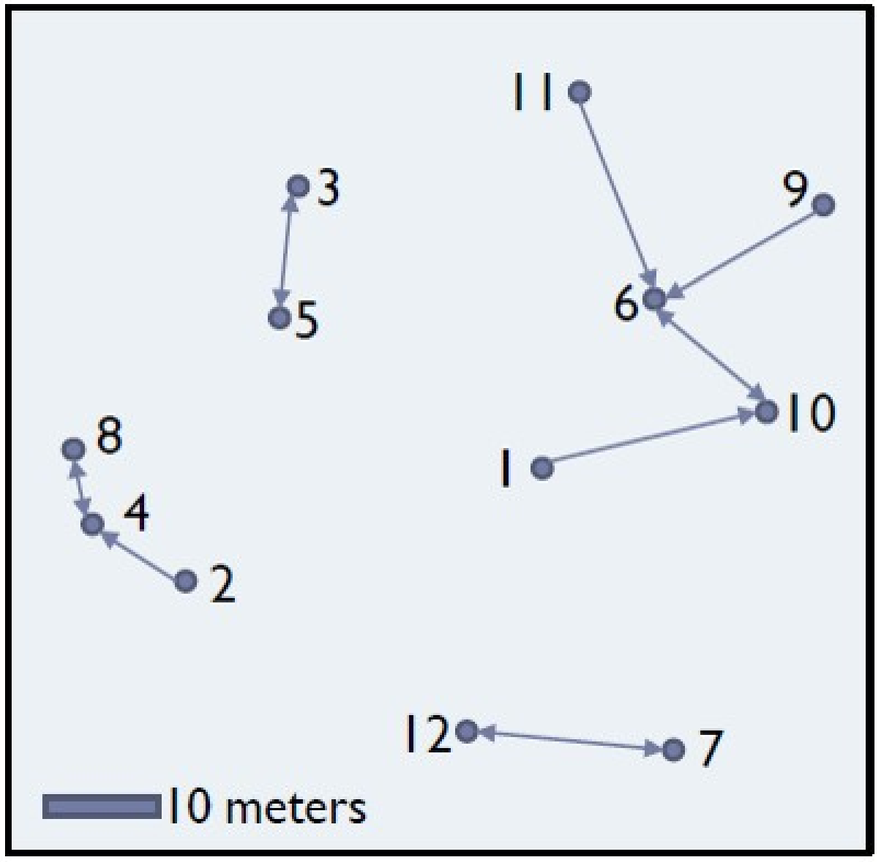
- It is sensitive to the quadrat size.
  - If the quadrat size is too small, they may contain only a couple of points, and
  - If the quadrat size is too large, they may contain too many points.
- It is a measure of **dispersion** rather than a measure of **pattern**.
- It results in a single measure for the entire distribution, so variation within the region are not recognised.



Distance-based: Nearest Neighbour Index

What is Nearest Neighbour?

Direct distance from a point to its nearest neighbour.



Event	x	y	Nearest neighbor	r_{min}
1	66.22	32.54	10	25.59
2	22.52	22.39	4	15.64
3	31.01	81.21	5	21.14
4	9.47	31.02	8	24.81
5	30.78	60.10	3	9.00
6	75.21	58.93	10	21.14
7	79.26	7.68	12	21.94
8	8.23	39.93	4	9.00
9	98.73	42.53	6	21.94
10	89.78	42.53	6	21.94
11	65.19	92.08	6	34.63
12	54.46	8.48	7	24.81

Nearest Neighbour Index

The Nearest Neighbour Index is expressed as the ratio of the **Observed Mean Distance** to the **Expected Mean Distance**.

NN Index: The Nearest Neighbor Index (Uncorrected)

$$NNI = \frac{\bar{d}}{E(\bar{d})}$$

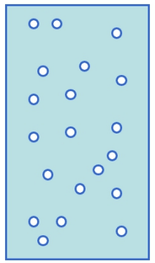
Avg. Dist.: Average Nearest Neighbor Distance

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

Exp. Avg.: Expected Average Nearest Neighbor Distance (Uncorrected)

$$E(\bar{d}) = 0.5\sqrt{\frac{A}{n}}$$

Calculating Nearest Neighbour Index

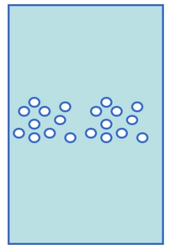


RANDOM

Point	Nearest Neighbor	Distance
1	2	1
2	3	0.1
3	2	0.1
4	5	1
5	4	1
6	5	2
7	6	2.7
8	10	1
9	10	1
10	9	1
		10.9

Mean distance	1.09
Area of Region	50
Density	0.2
Expected Mean	1.118034
NNI	0.974926

Z = -0.1515

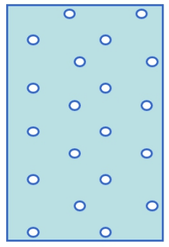


CLUSTERED

Point	Nearest Neighbor	Distance
1	2	0.1
2	3	0.1
3	2	0.1
4	5	0.1
5	4	0.1
6	5	0.1
7	6	0.1
8	9	0.1
9	10	0.1
10	9	0.1
		1

Mean distance	0.1
Area of Region	50
Density	0.2
Expected Mean	1.118034
NNI	0.089443

Z = 5.508



UNIFORM

Point	Nearest Neighbor	Distance
1	3	2.2
2	4	2.2
3	4	2.2
4	5	2.2
5	7	2.2
6	7	2.2
7	8	2.2
8	9	2.2
9	10	2.2
10	9	2.2
		22

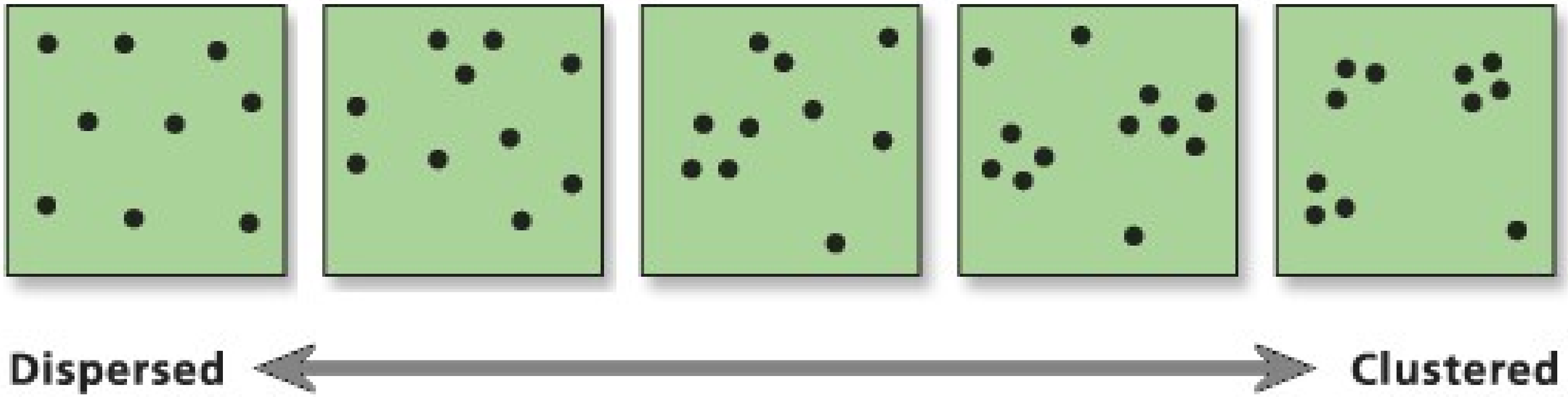
Mean distance	2.2
Area of Region	50
Density	0.2
Expected Mean	1.118034
NNI	1.96774

Z = 5.855

Interpreting Nearest Neighbour Index

The expected distance is the average distance between neighbours in a hypothetical random distribution.

- If the index is less than 1, the pattern exhibits clustering,
- If the index is equal to 1, the patterns exhibits random, and
- If the index is greater than 1, the trend is toward dispersion or competition.



The test statistics

- Null Hypothesis: Points are randomly distributed
- Test statistics:

$$Z = \frac{\bar{d} - E(\bar{d})}{Std.error}$$

- Reject the null hypothesis if the z-score is large and p-value is smaller than the alpha value.

Interpreting Nearest Neighbour Index

```
        Clark-Evans test
        No edge correction
        Monte Carlo test based on 999 simulations of CSR
        with fixed n

data:  childcareSG_ppp
R = 0.545, p-value = 0.002
alternative hypothesis: two-sided
```

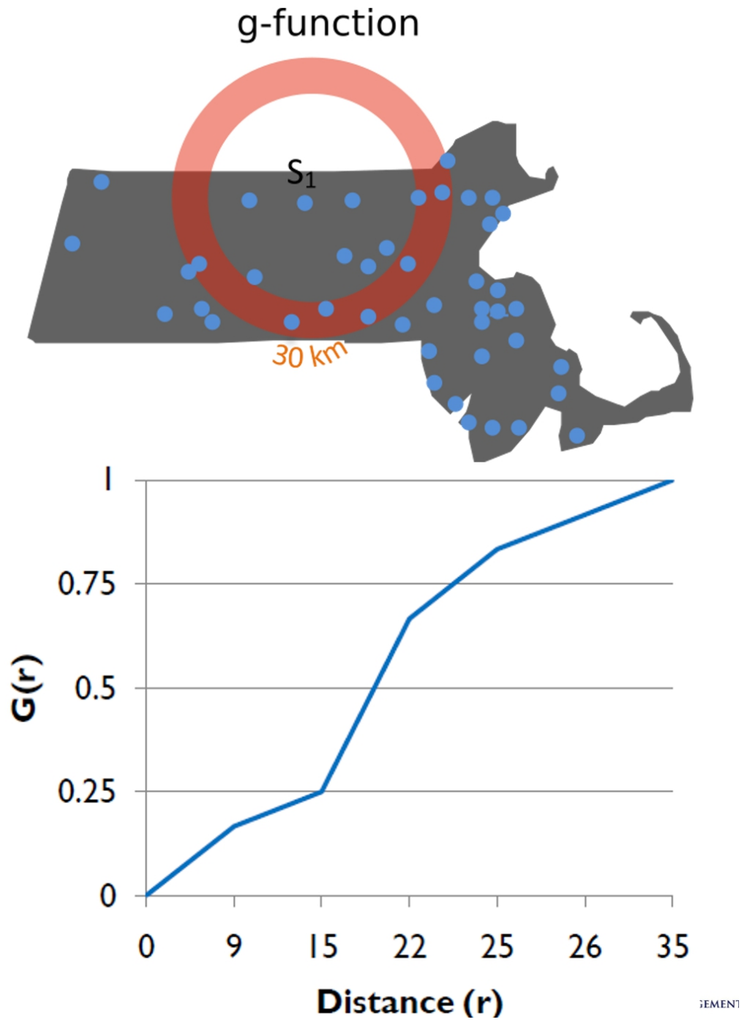
The p-value is smaller than 0.05 => Reject the null hypothesis that the point patterns are randomly distributed.

G function

The formula

$$G(r) = \frac{\#[r_{\min}(s_i) < r]}{n} = \frac{\#\,\text{point pairs where } r_{\min} \le r}{\#\,\text{of points in study area}}$$

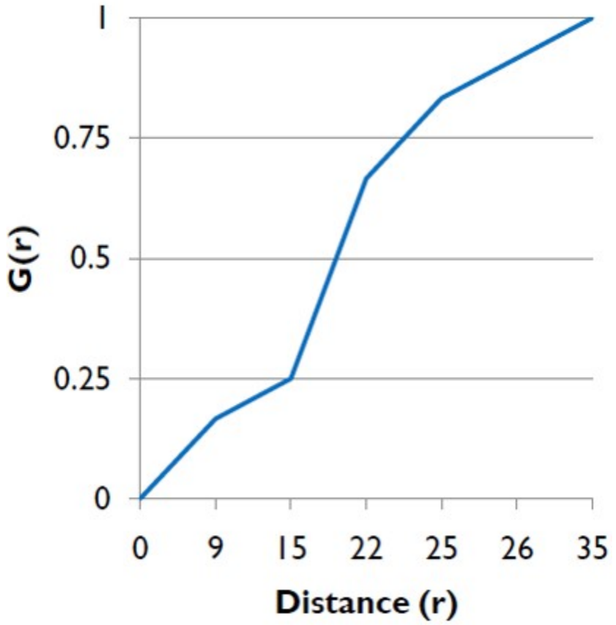
Event	x	y	Nearest neighbor	r_{min}
1	66.22	32.54	10	25.59
2	22.52	22.39	4	15.64
3	31.01	81.21	5	21.14
4	9.47	31.02	8	24.81
5	30.78	60.10	3	9.00
6	75.21	58.93	10	21.14
7	79.26	7.68	12	21.94
8	8.23	39.93	4	9.00
9	98.73	42.53	6	21.94
10	89.78	42.53	6	21.94
11	65.19	92.08	6	34.63
12	54.46	8.48	7	24.81



Interpretation of G-function

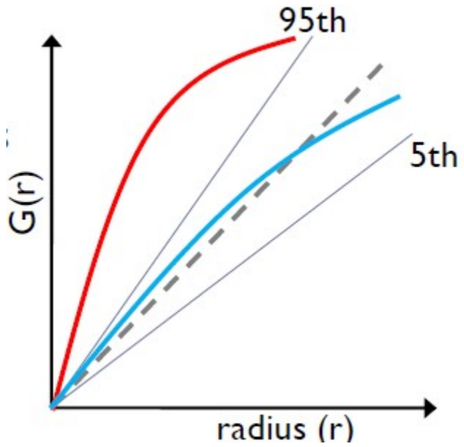
The shape of G-function tells us the way the events are spaced in a point pattern.

- Clustered: G increases rapidly at short distance.
- Evenness: G increases slowly up to distance where most events spaced, then increases rapidly.



How do we tell if G is significant?

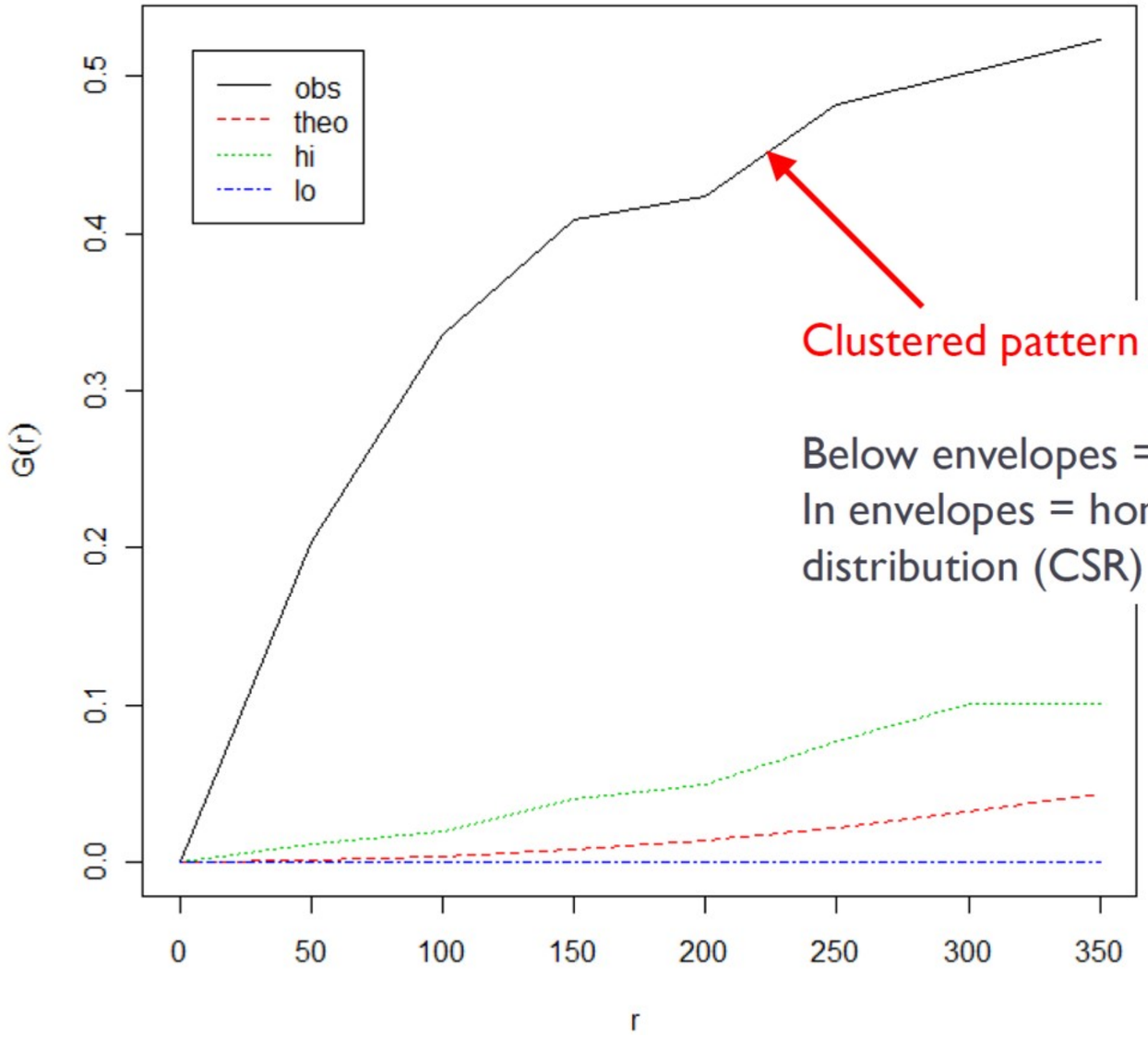
- The significant of any departure from CSR (either cluster or regularity) can be evaluated using simulated “confidence envelopes”



Monte Carlo simulation test of CSR

- Perform m independent simulation of n events (i.e. 999) in the study region.
- For each simulated point pattern, estimate G(r) and use the maximum (95th) and minimum (5th) of these functions for the simulated patterns to define an upper and lower simulation envelope.
- If the estimated G(r) lies above the upper envelope or below the lower envelope, the estimated G(r) is statistically significant.

The significant test of G-function



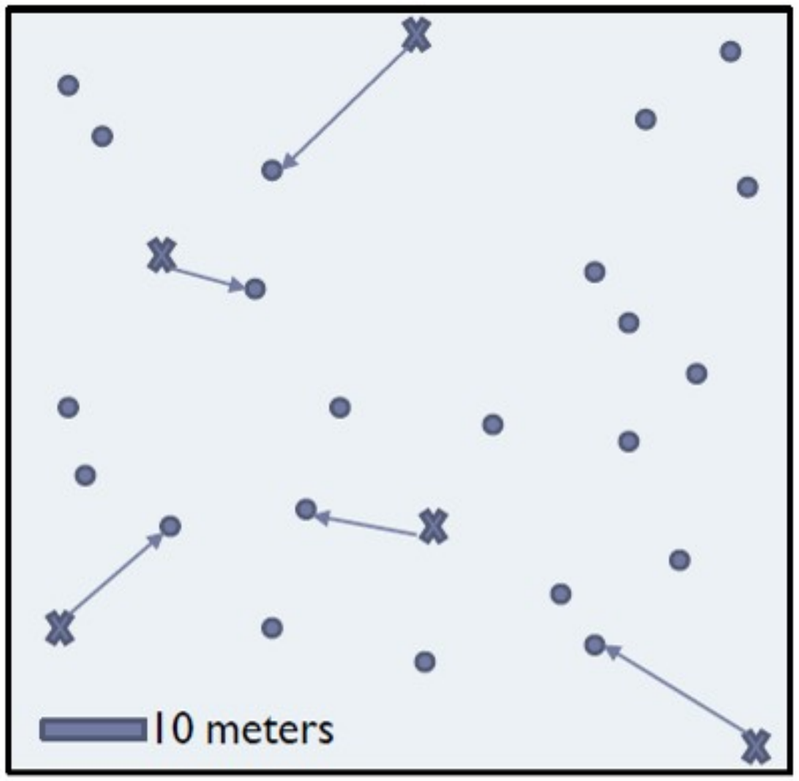
Clustered pattern (above the envelopes)

Below envelopes = regular pattern
In envelopes = homogeneous distribution (CSR)

F function

- Select a sample of point locations anywhere in the study region at random
  - Determine minimum distance from each point to any event in the study area.
- Three steps:
  - Randomly select m points (p1, p2, ….., pn),
  - Calculate dmin(pi,s) as the minimum distance from location pi to any event in the point patterns, and
  - Calculate F(d).

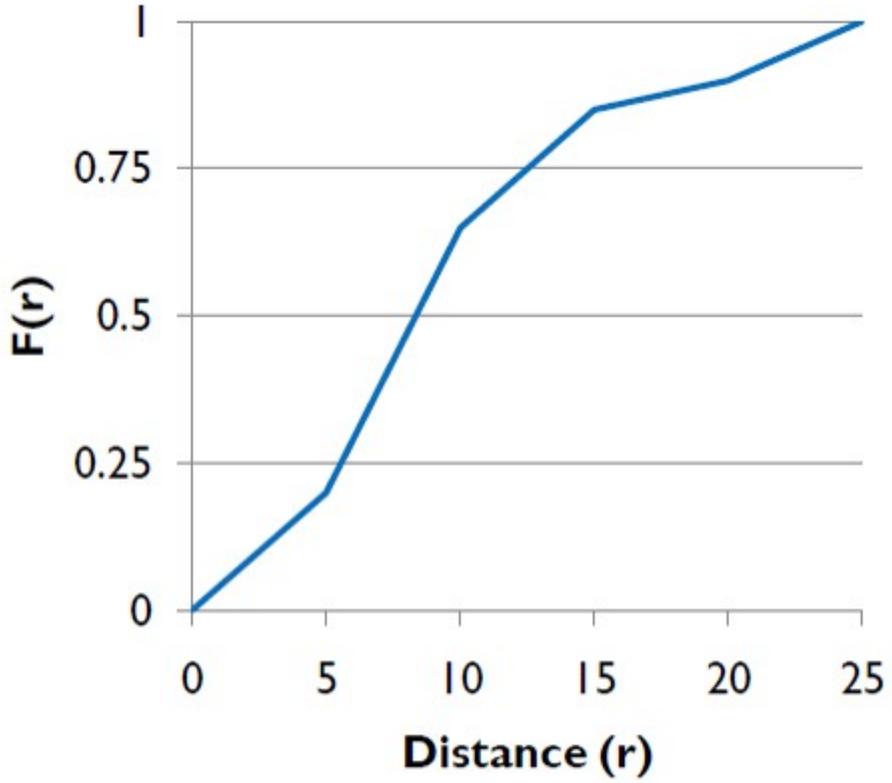
The F function formula



= randomly chosen point

= event in study area

= d_{min}

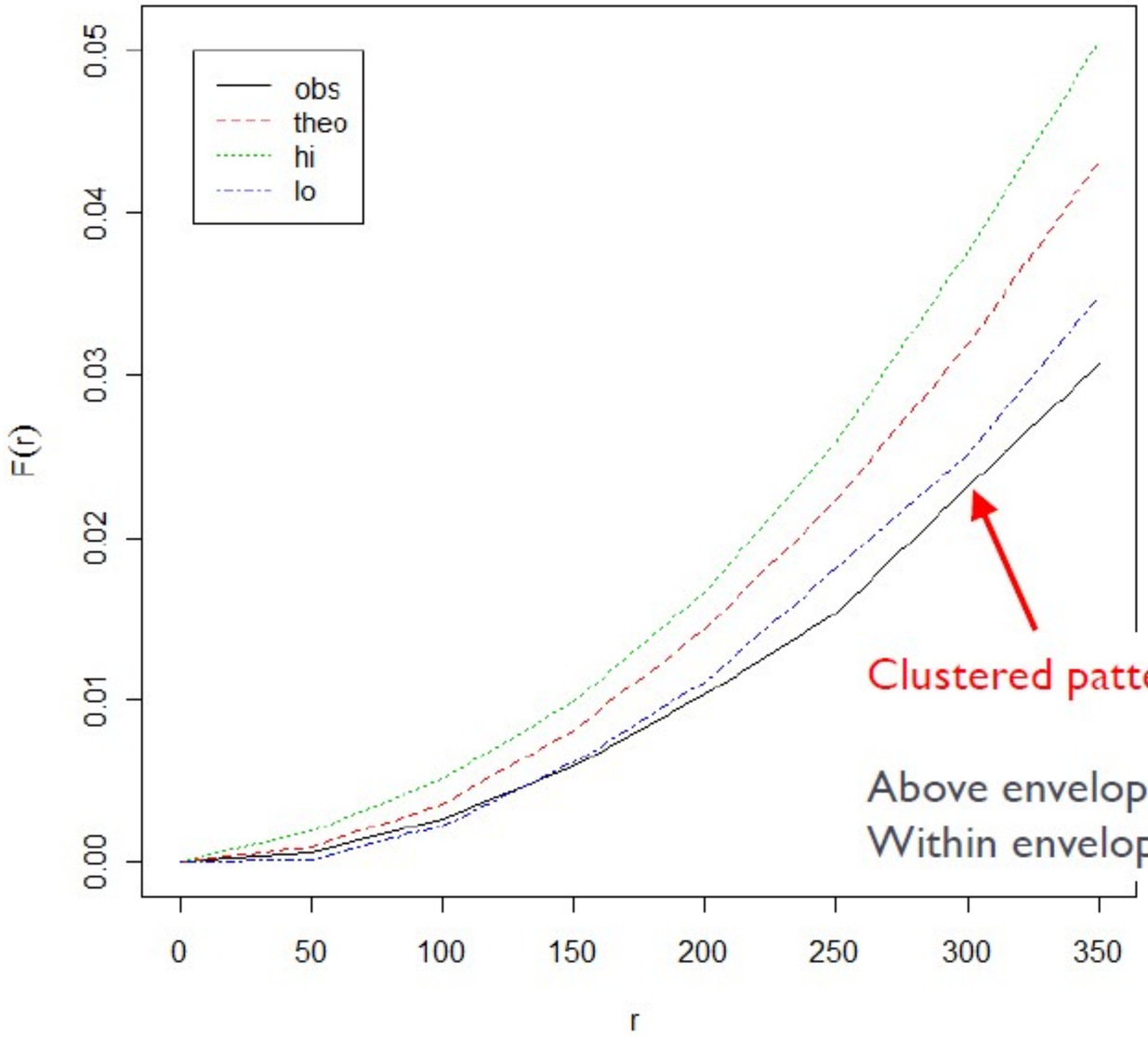


$$F(d) = \frac{\#[d_{\min}(p_i, s) < d]}{m} = \frac{\#\text{ of point pairs where } r_{min} \leq r}{\#\text{ sample points}}$$

Interpretation of F-function

- Clustered = F(r) rises slowly at first, but more rapidly at longer distances.
- Evenness = F(r) rises rapidly at first, then slowly at longer distances.

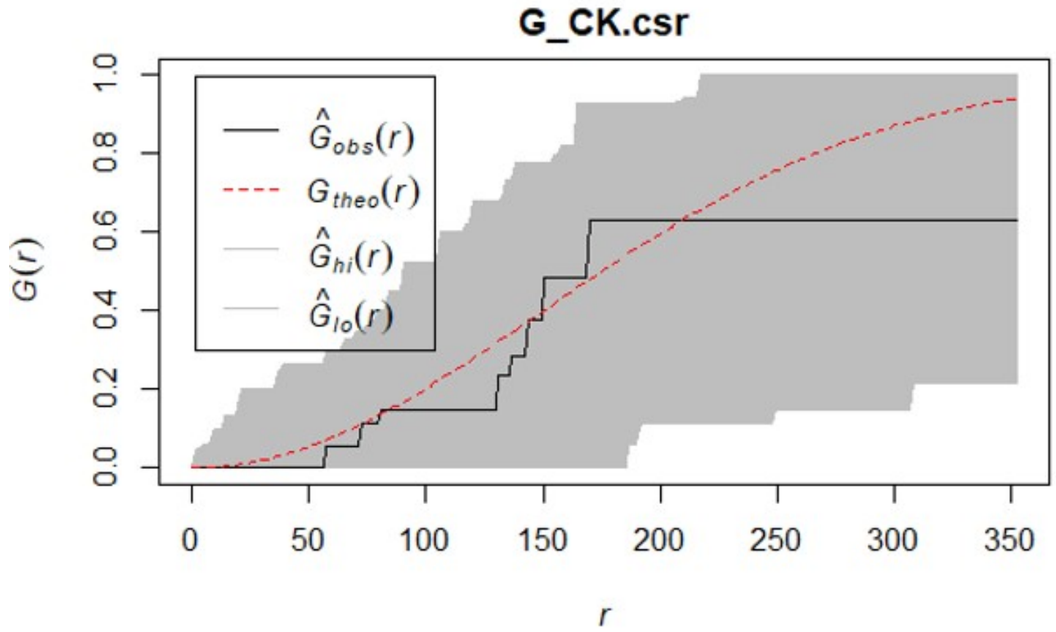
The significant test of F-function

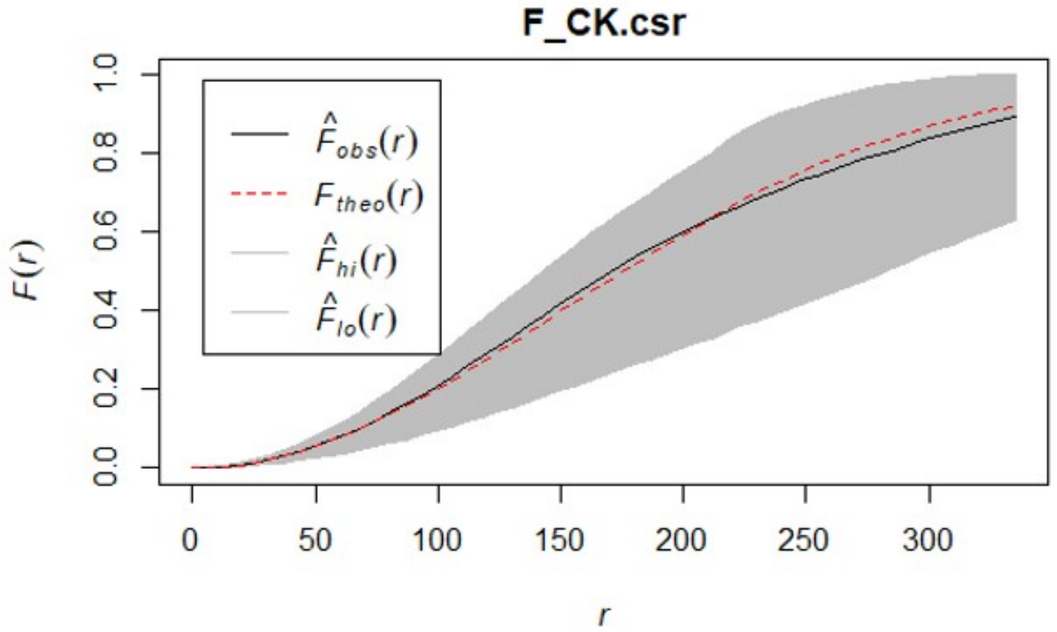


Clustered pattern (below the envelopes)

Above envelopes = regular pattern
Within envelopes = CSR

Comparison between G and F



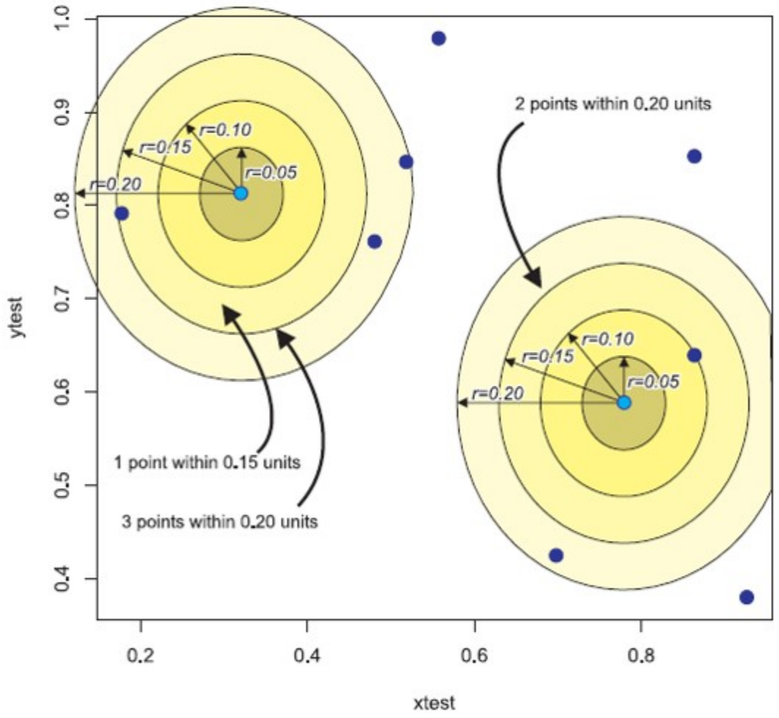


Ripley’s K function (Ripley, 1981)

- Limitation of nearest neighbor distance method is that it uses only nearest distance
- Considers only the shortest scales of variation.
- K function uses more points.
  - Provides an estimate of spatial dependence over a wider range of scales.
  - Based on all the distances between events in the study area.
  - Assumes isotropy over the region.

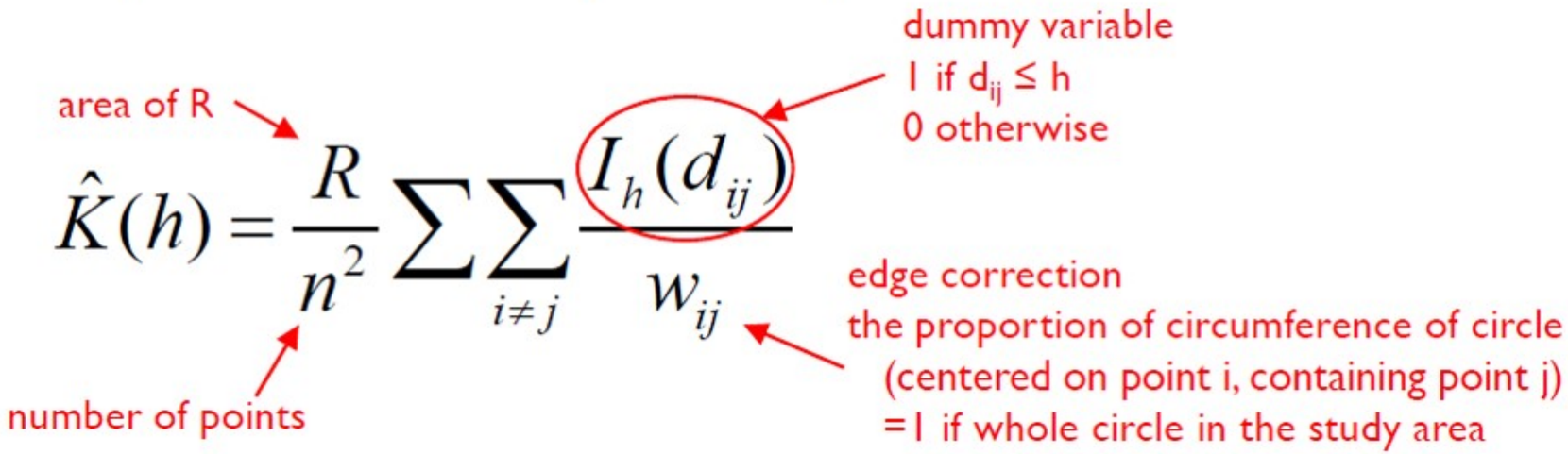
Calculating the K function

- Construct a circle of radius h around each point event(i).
- Count the number of other events (j) that fall inside this circle.
- Repeat these two steps for all points (i) and sum results.
- Increment h by a small amount and repeat the calculation.



K function

The formula:



The K function complete spatial randomness test

- K(h) can be plotted against different values of h.
- But what should K look like for no spatial dependence?
- Consider what K(h) should look like for a random point process (CSR)
  - The probability of an event at any point in R is independent of what other events have occurred and equally likely anywhere in R

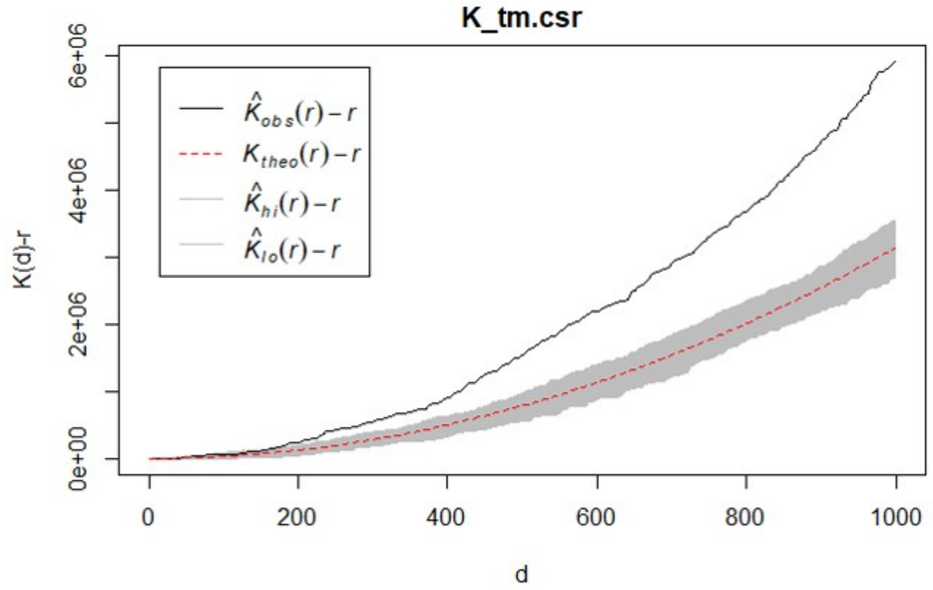
Interpreting the K function complete spatial randomness test

Under the assumption of CSR, the expected number of events within distance h of an event is:

$$K(h) = \pi h^2$$

where

h = the radius of the circle



Compare K(h) to πh^2

- K(h) < πh^2 if point pattern is regular
- K(h) > πh^2 if point pattern is clustered

- Above the envelop: significant cluster pattern - Below the envelop: significant regular
- Inside the envelop: CSR

The L function (Besag 1977)

In practice, K function will be normalised to obtained a benchmark of zero.

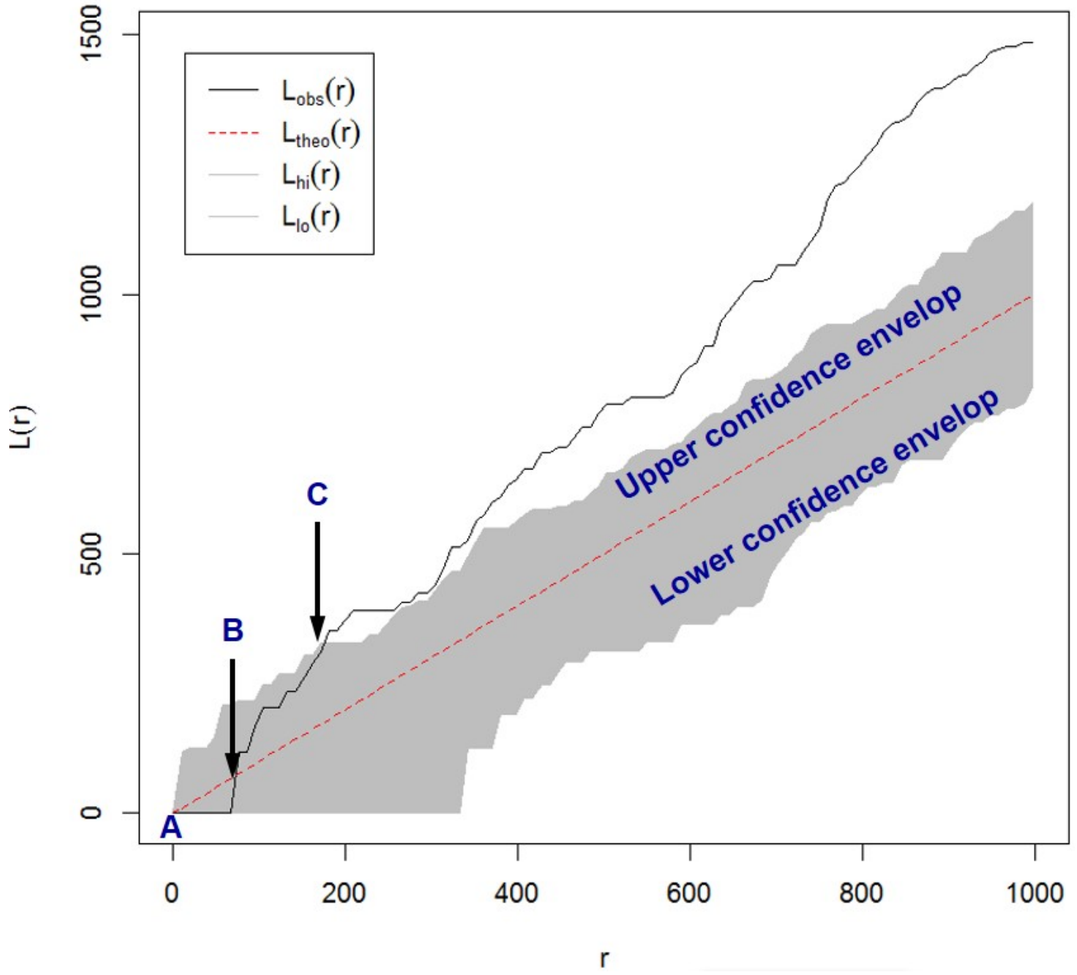
The formula:

$$L(r) = \sqrt{\frac{K(r)}{\pi}}.$$

Interpreting the L function complete spatial randomness test

- When an observed L value is greater than its corresponding L(theo)(i.e. red break line) value for a particular distance and above the upper confidence envelop, spatial clustering for that distance is statistically significant (e.g. distance beyond C).
- When an observed L value is greater than its corresponding L(theo) value for a particular distance and lower than the upper confidence envelop, spatial clustering for that distance is statistically NOT significant (e.g. distance between B and C).
- When an observed L value is smaller than its corresponding L(theo) value for a particular distance and beyond the lower confidence envelop, spatial dispersion for that distance is statistically significant. - When an observed L value is smaller than its corresponding L(theo) value for a particular distance and within the lower confidence envelop, spatial dispersion for that distance is statistically NOT significant (e.g. distance between A and B).

- The grey zone indicates the confident envelop (i.e. 95%).

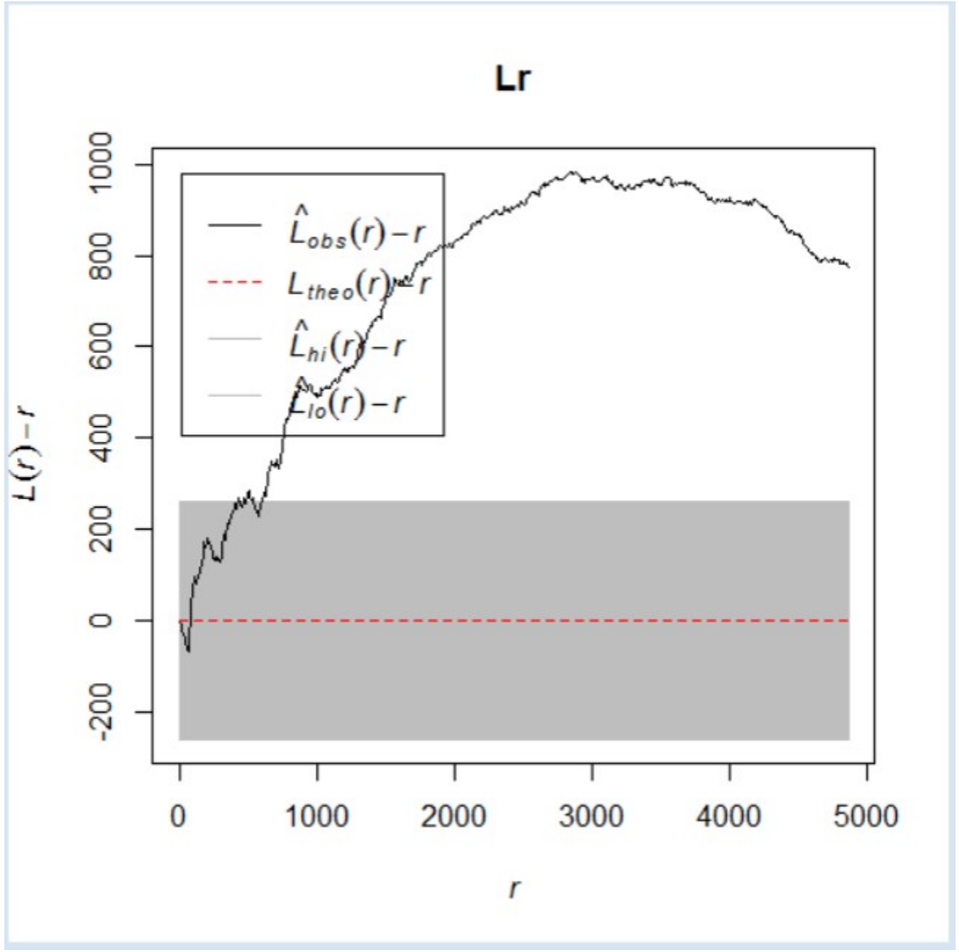


The L function (Besag 1977)

The modified L function

$$L(r) = \sqrt{\frac{K(r)}{\pi}} - r$$

- L(r)>0 indicates that the observed distribution is geographically concentrated.
- L(r)<0 implies dispersion.
- L(r)=0 indicates complete spatial randomness (CRS).



References

- Chapter 11 Point Pattern Analysis of **Intro to GIS and Spatial Analysis**. Section 11.2, 11.3, 11.3.1 and 11.4
- GIS&T Body of Knowledge AM-07-Point Pattern Analysis
- GIS&T Body of Knowledge AM-08-Kernels and Density Estimation
- Analyzing Patterns in Business Point Data, Directions Magazine March 17, 2005.
- O'Sullivan, D., and Unwin, D. (2010) **Geographic Information Analysis**, Second Edition. John Wiley & Sons Inc., New Jersey, Canada. Chapter 5-6.
- Baddeley A., Rubak E. and Turner R. (2015) **Spatial Point Patterns: Methodology and Applications with R**, Chapman and Hall/CRC.