Chapter 6
Spatial Analysis

6.1 Introduction

Spatial analysis, in a narrow sense, is a set of mathematical (and usually statistical) tools used to find order and patterns in spatial phenomena.

Spatial patterns found in spatial analysis help our understanding of not only spatial phenomena themselves but also their underlying structure, because spatial patterns are realization of interactions between spatial objects.

Once we understand the underlying structure of a spatial phenomenon, we can describe it using a mathematical framework - this is what we call 'spatial modeling'. Using spatial models we can simulate spatial phenomena in computer environment.

Spatial analysis leads to spatial modelling and simulation.

What is spatial pattern?

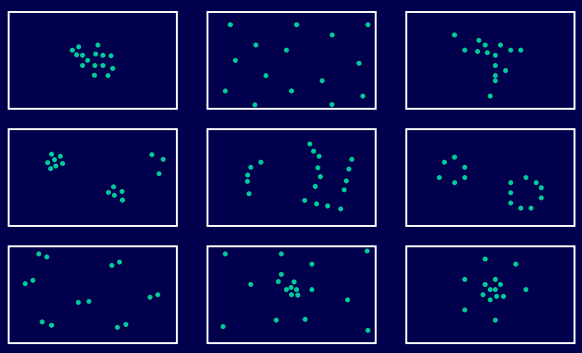


Figure: Patterns of points

The term '**spatial pattern**' is not firmly defined in spatial analysis. It often refers to various levels of spatial regularity, which includes local clusters of points, global structure of a surface, and so forth.

Some spatial patterns can be easily found by visual analysis, while others are so complicated that it is difficult to detect them only by visual analysis.

Objectives of spatial analysis are

1. to detect spatial patterns that cannot be detected by visual analysis, and
2. to confirm whether a spatial pattern found in visual analysis is significant (nonrandom).

To achieve the first objective, we do '**exploratory spatial analysis**', which is considered as a part of '**spatial data mining**'.

The second objective emphasizes statistical analysis, because it borrows statistical techniques from spatial statistics. It is often called '**confirmatory spatial analysis**'.

Spatial analysis started as a subject-dependent statistics in biometrics, epidemiology, and econometrics. Then spatial analysis drew attention of statisticians in 1970's, and the methodology was drastically sophisticated.

Though spatial analysis includes non-statistical analysis, it is often called '**spatial statistics**' because a great part of spatial analysis is based on statistical analysis.

One of the questions frequently raised in spatial analysis is

'Is this spatial pattern statistically significant?'

We want to know whether a pattern emerged only by chance or it appears from certain causes.

• References - Overview of spatial analysis

1. Fotheringham, A. S. and Rogerson, P. (1993): *Spatial Analysis and GIS*, Taylor and Francis.
2. Bailey, B. C. and Gatrell, A. C. (1995): *Interactive Spatial Data Analysis*, Prentice-Hall.
3. O'Sullivan, D. and Unwin, D. (2002): *Geographic Information Analysis*, John Wiley.
4. Wong, D. W.-S. and Lee, J. (2005): *Statistical Analysis with ArcView GIS and ArcGIS*, John Wiley.

5. Longley P. and Batty, M. (2003): *Advanced Spatial Analysis: The CASA book of GIS*, ESRI Press.
6. Maguire, D. J., Batty, M., and Goodchild, M. F. (2005): *GIS, Spatial Analysis and Modeling*, ESRI Press.

• References – General Statistics

1. Hoel, A. C. (1984): *Introduction to Mathematical Statistics, 5th Edition*, John Wiley.
2. Wonnacott, T. H. and Wonnacott, R. J. (1990): *Introductory Statistics, 5th Edition*, Jown Wiley.
3. Wonnacott, T. H. and Wonnacott, R. J. (1990): *Introductory Statistics for Business and Economics, 4th Edition*, Jown Wiley.
4. Fox, J. (1997): *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publication.
5. Chatterjee, S., Price, B., and Hadi, A. S. (1999): *Regression Analysis by Example, 3rd Edition*, John Wiley.

• References – Spatial statistics

1. Ripley, B. D. (1981): *Spatial Statistics*, John Wiley.
2. Upton, G. and Fingleton, B. (1985): *Spatial Data Analysis by Example: Volume 1: Point Pattern and Qualitative Data*, John Wiley.
3. Cressie, N. (1993): *Statistics for Spatial Data, 2nd Edition*, John Wiley.
4. Moore, M. (2001): *Spatial Statistics : Methodological Aspects and Applications*, Springer.

• References – Spatial statistics

5. Diggle, P. (2002): *Statistical Analysis of Spatial Point Patterns*, Oxford University Press.
6. Schabenberger, O. and Gotway, C. A. (2004): *Statistical Methods For Spatial Data Analysis* , Chapman & Hall.

6.2 Point pattern analysis 1: homogeneous points

Points are the most fundamental spatial objectsin GIS. They are used for representing zero-dimensional spatial objects, that is, locations in a two- or higher dimensional space.

In GIS, however, points are also used for representing spatial objects including lines and polygons that are relatively smaller than the study region. The distribution of retail stores, for example, is represented as a point distribution, though stores are at least two-dimensional spatial objects in the real world.

Analysis of point distributions, which is often called 'point pattern analysis', is one of the basic methods in spatial analysis.

Point pattern analysis is applicable to any spatial distribution represented as a set of points in GIS.

• References - Point pattern analysis

1. Ripley, B. D. (1981): *Spatial Statistics*, John Wiley.
2. Upton, G. and Fingleton, B. (1985): *Spatial Data Analysis by Example: Volume 1: Point Pattern and Qualitative Data*, John Wiley.
3. Cressie, N. (1993): *Statistics for Spatial Data, 2nd Edition*, John Wiley.

Point pattern analysis, in its basic form, deals with the distribution of homogeneous points, that is, one type of points.

This does not imply that we cannot treat points that are not homogeneous in the real world. In basic point pattern analysis, we focus only on the spatial aspect of point distributions, neglecting their attributes.

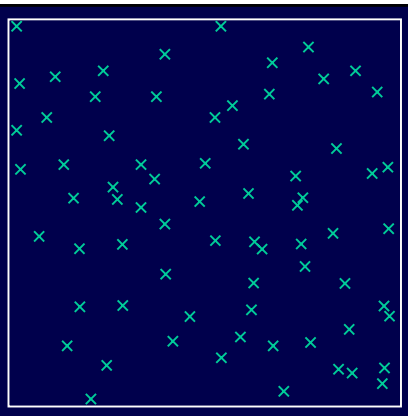


Figure: Distribution of Swedish Pine saplings (10m × 10m)

In point pattern analysis, it is important to consider whether a point distribution shows a clustered pattern or dispersed pattern.

If points represent cases of a disease, a point cluster suggests that the disease is epidemic or that there is a source of water pollution near the cluster.

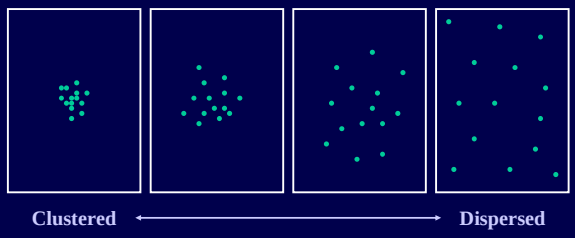


Figure: Point distributions

Because of this, in point pattern analysis we use a quantitative measure that indicates the degree of clustering.

6.2.1 Nearest neighbor distance method

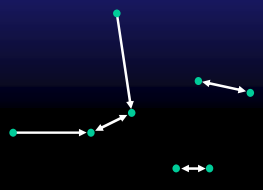
To describe the degree of spatial clustering of a point distribution, nearest neighbor distance method uses the average distance from every point to its nearest neighbor point.

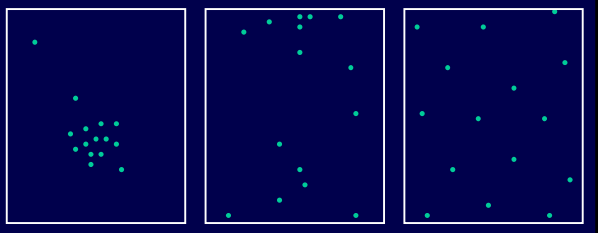
d_i: Distance to the nearest point from point i

n: Number of points

W: (Average) nearest neighbor distance

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





W=23.45 W=35.71 W=72.85

Figure: Nearest neighbor distance

The nearest neighbor distance defined above is an 'absolute' measure of point clusters. It depends on the size of the region in which points are distributed, so we cannot compare two sets of points distributed in regions of different sizes.

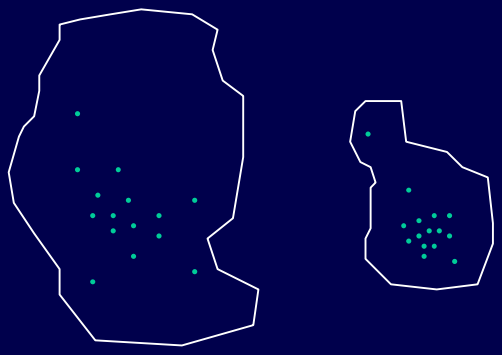


Figure: Which distribution is more clustered?

This indicates that the concept of spatial cluster is based on the pattern of points with respect to the size of region in which the points are located.

To evaluate the degree of spatial clustering, therefore, we have to standardize the nearest neighbor distance, taking the region size into account.

Point distribution under CSR

To evaluate a point distribution, we consider the point distribution under CSR (Complete Spatial Randomness). Points are distributed randomly over an infinite space.

This type of point distribution is called a homogeneous Poisson distribution. Homogeneous Poisson distribution has one parameter λ, the density of points.

The expectation of the nearest neighbor distance of points under CSR is represented as a function of point density λ:

$$\mathrm{E}[W] = \frac{1}{2\sqrt{\lambda}}$$

We can standardize the nearest neighbor distance *W* by dividing it by its expectation under CSR.

S: The region in which points are distributed

A: The area of *S*

The point density in *S* substitutes for the point density λ:

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

This is an approximate calculation, because homogeneous Poisson distribution is defined only in an infinite space.

Standardized nearest neighbor distance

The standardized nearest neighbor distance is then given by

$$w = \frac{W}{\mathrm{E}[W]}$$

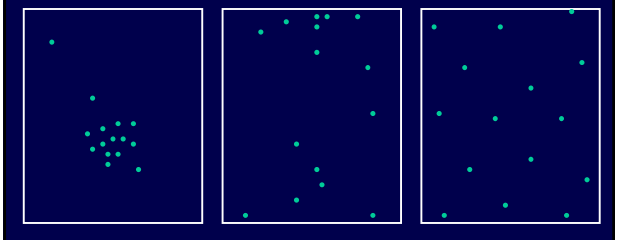
$$= 2W\sqrt{\frac{n}{A}}$$

Small *w* indicates spatial clustering of points, while large value indicates points are dispersed.

w < 1: Points are clustered

w ≈ 1: Points are randomly distributed

w > 1: Points are dispersed



w=0.5932 *w*=0.9034 *w*=1.8429

Figure: Standardized nearest neighbor distance

Statistical test

The standardized nearest neighbor distance is a descriptive measure of the degree of point clustering. Once we calculate it for a point distribution and obtain a small value, we want to know whether we can say with confidence "points are clustered."

To answer this question, we consider whether such a spatial clustering rarely occurs when points are randomly distributed, or it frequently happens and thus it is not significant. This is a general procedure of statistical tests.

Hypotheses

In statistical tests, we build null and alternative hypotheses.

Null hypothesis H_0:

Points are randomly distributed, following a homogeneous Poisson distribution.

Alternative hypothesis H_1:

Points are spatially clustered (or dispersed).

In significance test, we use the nearest neighbor distance W as a statistic.

If W is significantly small (large), we accept the alternative hypothesis H_1, and we can say that the points are clustered (dispersed) at a certain significance level. Otherwise, accepting the null hypothesis H_0, we say that the points are randomly distributed.

For statistical test, we need the probability distribution of the statistic W under the null hypothesis, CSR.

If the points are randomly distributed over an infinite space, the probability distribution of W is given by a normal distribution:

$$N\left(\frac{1}{2\sqrt{\lambda}}, \frac{4-\pi}{4\pi n\lambda}\right)$$

In reality, however, points are distributed in a bounded region, which makes the probability distribution of W under CSR slightly different from the normal distribution.

This is called '**edge effect**', which has to be corrected in the statistical test.

Correction of the edge effect depends on the number of points.

1) When n is large enough (>100), we randomly choose m sample points and calculate the average nearest neighbor distance. We can excludes the edge effect by the random sampling.

$$W = \frac{1}{m}\sum_{i=1}^{m} d_i$$

The average nearest neighbor distance follows a normal distribution

$$N\left(\frac{1}{2\sqrt{\lambda}}, \frac{4-\pi}{4\pi m\lambda}\right)$$

Consequently,

$$z = \frac{W - \frac{1}{2\sqrt{\lambda}}}{\sqrt{\frac{4-\pi}{4\pi m\lambda}}}$$

follows the standard normal distribution when points are randomly distributed.

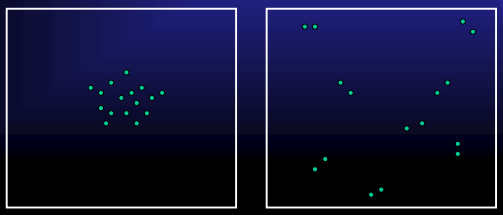
2) When n is not so large (n<100), we use all the n points in calculation of the nearest neighbor distance. In this case, the probability distribution of W under CSR is approximated by a normal distribution

$$N\left(0.5\sqrt{\frac{A}{n}} + 0.051\frac{L}{n} + 0.041\frac{L}{n\sqrt{n}},\ 0.070\frac{A}{n^2} + 0.037\sqrt{\frac{A}{n^5}}\right)$$

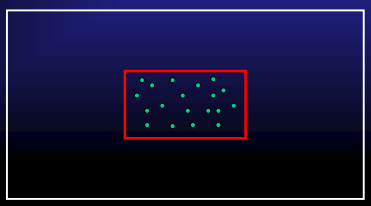
L: Perimeter of S

Limitations of the nearest neighbor distance method

1. We cannot distinguish all point distributions only by the nearest neighbor distance.



2. The result depends on the definition of S, the region in which points are distributed.



6.2.2 *K*-function method

K-function method overcomes the first limitation of the nearest neighbor distance method. *K*-function method can distinguish various types of point distributions not distinguishable by the nearest neighbor distance method.

Definition of *K*-function

K-function indicates the average number of points within a certain distance *h* from points. It is, therefore, a function of the distance *h*.

K-function counts the number of points located in the circular region of radius *h* from every point in *S*, and divides it by the number and density of points for standardization.

In contrast to the nearest neighbor distance, the *K*-function shows a large value when points are clustered.

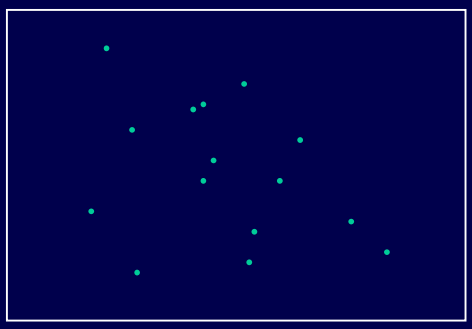


Figure: A point distribution

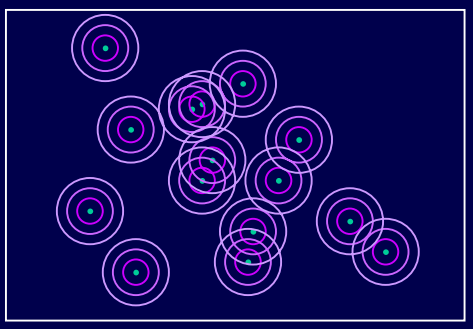


Figure: Calculation of *K*-function

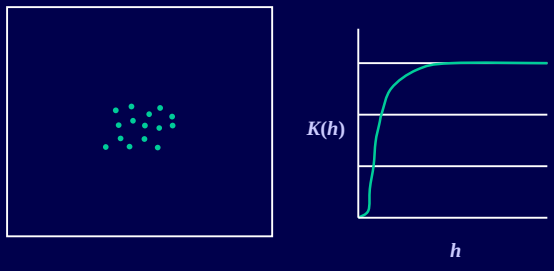


Figure: Example of *K*-function

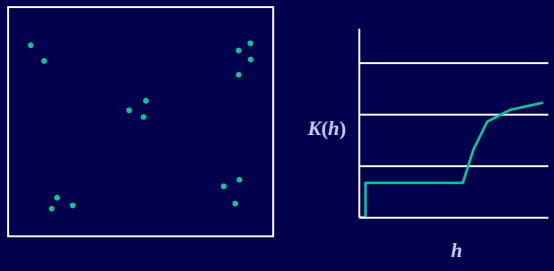


Figure: Example of *K*-function

Formal definition of *K*-function

***n*:** Number of points
x_i: Locational vector of point ***i***
λ: Point density (=n/A, ***A*:** the area of the region)
***h*:** Distance parameter
σ_{ij}: Binary function defined by

$$\sigma_{ij}(h) = \begin{cases} 1 & if \ |\mathbf{x}_i - \mathbf{x}_j| \le h \\ 0 & \text{otherwise} \end{cases}$$

***K*-function is mathematically defined by**

$$K(h) = \frac{\sum_i \sum_{j \neq i} \sigma_{ij}(h)}{n\lambda}$$

The numerator is the total number of points within a certain distance *h* from points.

***K*-function describes the degree of spatial clustering at the scale represented by the distance parameter *h*.**

A large *h* implies that we are discussing the point distribution at a small scale, in other words, in a large spatial extent. If *K*(*h*) shows a large value for a large *h*, we say that points are globally clustered, or to be exact, points are clustered at the scale of *h*.

Since both the *K*-function and the nearest neighbor distance methods use distance between points, they are often called 'distance methods**'.**

K-function of points under CSR

To evaluate the degree of spatial clustering of points, we again consider the distribution of points under CSR.

The expectation of *K*-function of points under CSR is

$$\mathrm{E}[K(h)] = \pi h^2$$

Comparing *K*(*h*) and its expectation, we can classify point distributions into one of three categories:

***K*(*h*) > πh^2:** Points are clustered
***K*(*h*) ≈ πh^2:** Points are randomly distributed
***K*(*h*) < πh^2:** Points are dispersed

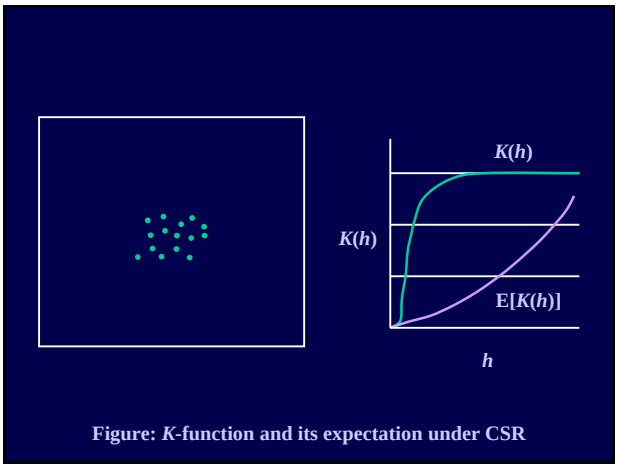


Figure: *K*-function and its expectation under CSR

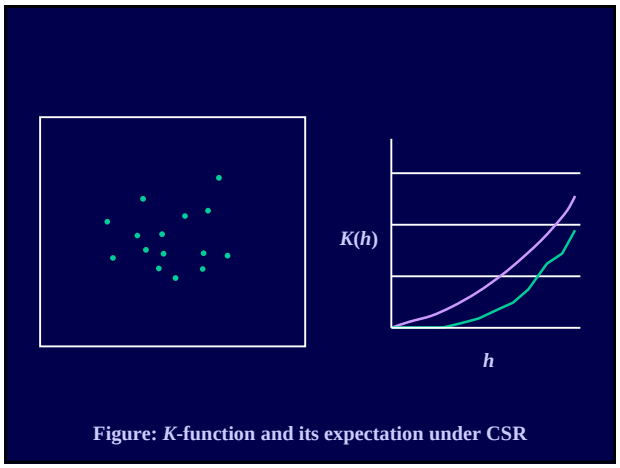


Figure: *K*-function and its expectation under CSR

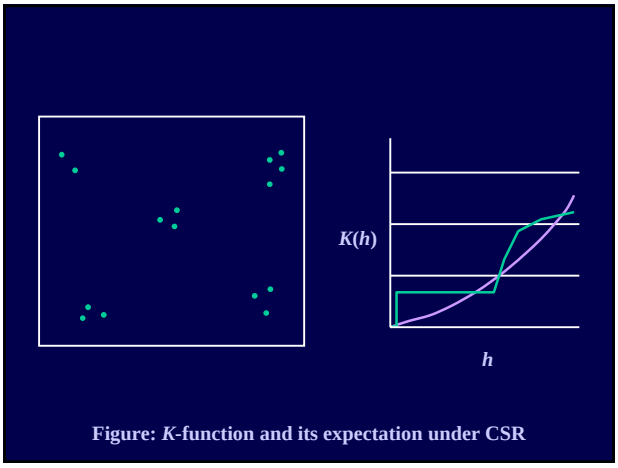


Figure: *K*-function and its expectation under CSR

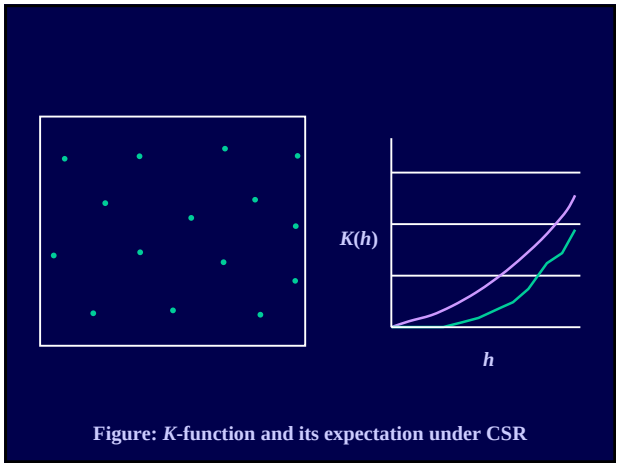


Figure: *K*-function and its expectation under CSR

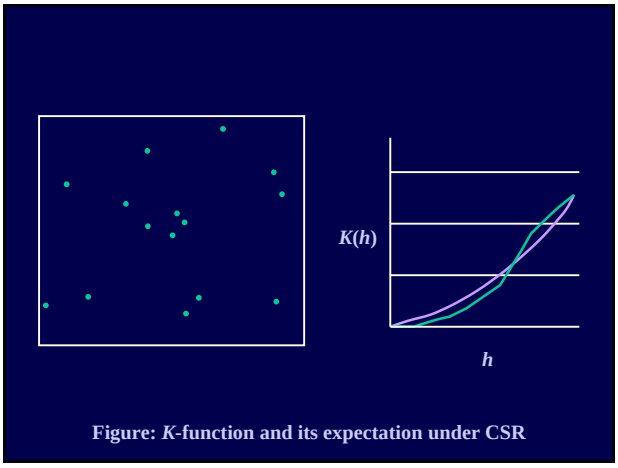


Figure: *K*-function and its expectation under CSR

Standardization of *K*-function

To compare *K*-function values among different *h* values, we standardize the *K*-function

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

The standardized *K*-function is called *L*-function.

Point distributions are then classified as below.

$L(h) > 0$: Points are clustered
$L(h) \approx 0$: Points are randomly distributed
$L(h) < 0$: Points are dispersed

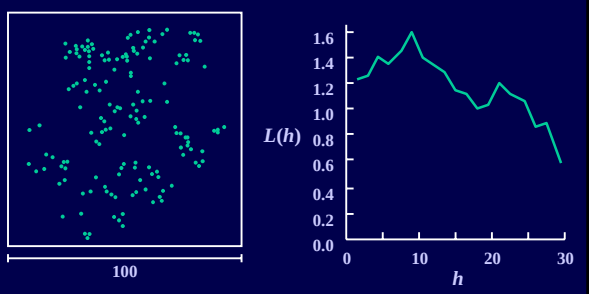


Figure: Juvenile offenders in Cardiff

Statistical test

Statistical test is applicable also to the *K*-function method. We can answer questions such as "whether points are significantly clustered?"

Hypotheses

Null hypothesis H_0:
Points are randomly distributed, following a homogeneous Poisson distribution.

Alternative hypothesis H_1:
Points are spatially clustered (or dispersed).

If $K(h)$ is significantly large (small), we accept the alternative hypothesis H_1, and we can say that the points are clustered (dispersed). Otherwise, accepting the null hypothesis H_0, we say that the points are randomly distributed.

For statistical test we need the probability distribution of $K(h)$ of points under CSR, which depends on the number of points n.

1) When n is large enough (n>100), we randomly choose m sample points and calculate the K-function. The probability distribution of the K-function of points under CSR is approximately given by a normal distribution

$$N\left(\pi h^2, \frac{\pi h^2}{m\lambda}\right)$$

2) When n is not so large (n<100), the probability distribution of points under CSR cannot be represented in an analytical form. In such a case, we often use Monte Carlo simulation. We repeatedly distribute n points randomly in S and calculate the K-function. Repeating this at least 10,000 times, we can obtain an approximate probability distribution of points under CSR.

6.2.3 Quadrat method

Quadrat method, which is often used in visual analysis, can also be used for statistical test.

Quadrat method converts point data into raster data by counting the number of points in every cell, and performs statistical test.

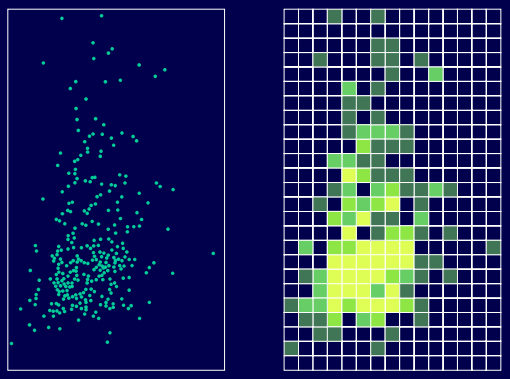


Figure: Quadrat method

Hypotheses

The quadrat method differs from the distance methods in statistical hypotheses.

Null hypothesis H_0:
Points are distributed according to the uniform distribution in S. This hypothesis is equivalent to the dispersed distribution discussed in the distance methods.

Alternative hypothesis H_1:
Points are spatially clustered.

c: Number of cells covering S.
x_i: Number of points in cell i.
$\overline{x}$: Average number of points in a cell

The quadrat method uses the χ^2 statistic defined by

$$\chi^2 = \frac{\sum_{i=1}^{c}\left(x_i - \overline{x}\right)^2}{\overline{x}}$$

If points are uniformly (dispersedly) distributed, the χ^2 statistic shows a small value, because x_is will be close to the mean $\overline{x}$. If χ^2 is very small, we accept the null hypothesis and say that points are uniformly distributed.

If points are spatially clustered, the χ^2 statistic shows a large value. If χ^2 is large enough, we can reject the null hypothesis and conclude that points are not uniformly distributed.

When points are distributed according to the uniform distribution, the χ^2 statistic approximately follows the χ^2 distribution with c degrees of freedom.

Consequently, we can test the significance of χ^2 by the ordinary χ^2 test which is used in basic statistics.

χ^2 test

The χ^2 test is widely used in statistics. Basically, it is a test for comparing an observed sample distribution with a distribution derived from a theoretical model.

The χ^2 statistic, in its original form, is

$$\chi^2 = \frac{\sum_{i=1}^{c} (x_i - y_i)^2}{y_i}$$

where y_i is the expectation of x_i when events follow the theoretical distribution. The χ^2 statistic shows a small value if the theory fits the observed data.

In point pattern analysis, the theoretical distribution considered in the χ^2 test is the uniform distribution in S. The expectation of x_i, which is denoted by y_i, is thus given by the average number of points in a cell:

$$\begin{aligned} y_i &= \frac{1}{n}\sum_{i=1}^{c} x_i \\ &= \overline{x} \\ &= \frac{n}{c} \end{aligned}$$

History

The quadrat method is first used in geography by a Japanese geographer Isamu Matsui.

Matui, I. (1932): Statistical study of the distribution of scattered villages in two regions of the Tonami Plain, Toyama Prefecture, *Japanese Journal of Geology and Geography*, 9, 251-255.

Advantage of quadrat method

One of the advantages of the quadrat method is that we can analyze a point distribution statistically in comparison of any distribution derived from a theory.

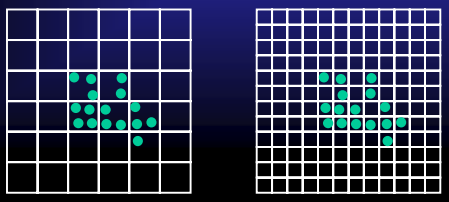
In the null hypothesis, we can consider not only the uniform distribution but also any other distribution derived from a spatial model, say, inhomogeneous Poisson distribution, a clustered distribution, etc..

Limitations of quadrat method

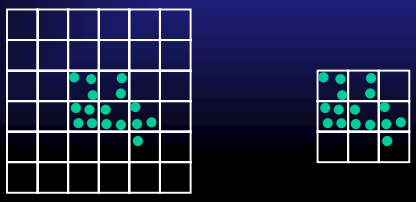
The quadrat method aggregates point data into raster data. This implies that the quadrat method ignores a large amount of locational information in the observed point distribution.

Because of this, the quadrat method has several limitations to which we should pay attention.

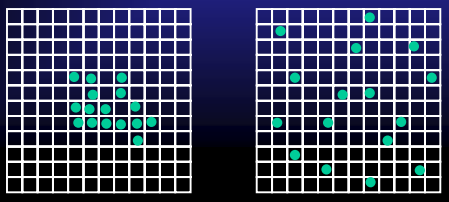
1. The result depends on the cell size.



2. The result depends on the definition of the region in which points are distributed.



3. The quadrat method cannot distinguish some different distributions.



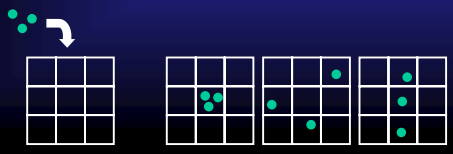
Those limitations are quite similar to those of the nearest neighbor distance method. Consequently, one solution is to try various cell size and interpret the result as a function of the spatial scale represented by the cell size. This method is parallel to the *K*-function method.

Even if we do so, if cells contain only a few points, statistical test does not work successfully. We cannot reject the null hypothesis and always have to say "points are uniformly distributed."

To solve this problem, we should use large cells each of which contain five or more points. This reduces the number of empty cells, which leads to a meaningful statistical test.

Homework Q.6.1

Suppose a 3 x 3 square lattice as shown below. We locate three points randomly on the lattice. It is prohibited to locate the points on the lattice boundary.

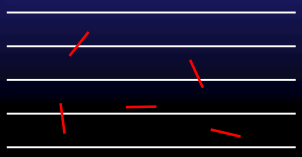


Homework Q.6.1 (cntd.)

1. What is the probability that all the points are located in the same cell?
2. What is the probability that the three points are arranged lengthways?
3. What is the probability that the three points are arranged lengthways, breadthways, or diagonally, as seen in the bingo game winner?

Homework Q.6.2

Suppose parallel lines equally spaced by a distance l on an infinite plane. We randomly drop a line segment of length l. What is the probability that the line segment crosses one of the parallel lines?



Homework Q.6.3

Suppose a square lattice of interval d. We randomly drop a circle of radius r ($r>d$). What is the expected number of lattice points contained in the circle?