Geo Factsheet

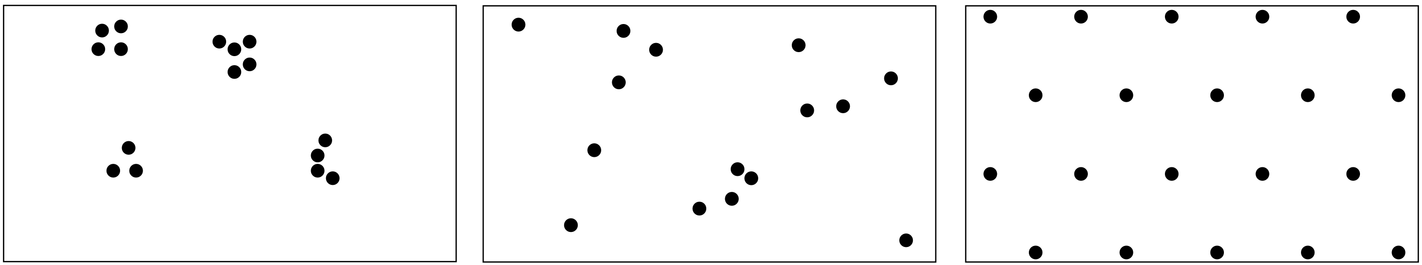
www.curriculum-press.co.uk Number 168

Nearest Neighbour Index

The Nearest Neighbour Index is a method of assessing the **spatial distribution** of points - in other words, how spread out they are. This Factsheet will

- introduce the two types of nearest neighbour index and discuss when to use each;
- show how to calculate the indices, test their significance and interpret the results;
- highlight some common pitfalls and limitations;
- give some suggestions for ways to use Nearest Neigbour.

The Nearest Neighbour Index measures whether points are:



clustered **random** **regularly dispersed**

The points can be villages, shops, people on a beach, trees in a woodland...

The index takes values between 0 and 2.15:

0	1	2.15
clustered	**random**	**regular dispersed**

← *increasingly clustered* *increasingly dispersed* →

Types of index

The index looks at the distances between each point and the closest other point – its ***nearest neighbour.***

There are two variants of the nearest neighbour index:

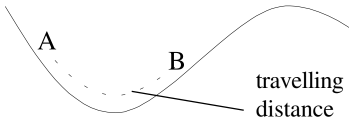
- **Spatial** - for points spread out in an area
  eg trees in a wood, shoe shops in a town centre
- **Linear -** for points on a line
  eg shoe shops along the single main shopping street, car accidents along a stretch of road

When should I use linear?

When the points (shops, accidents etc...) can only locate along a specific line, rather than anywhere in space, then the linear version should be used.

The line does not have to be straight - it can be curved, or consist of two lines at an angle, like some streets.

NB: When using the linear version, measure the distances **along the line** (i.e. along the road), rather than "as the crow flies". The **travelling** distance between the points is the important one here.



Calculating the index

The formulae are:

Spatial

$$R = 2\overline{D}\sqrt{\frac{n}{A}}$$

Linear

$$R = 2\overline{D}\left(\frac{n-1}{L}\right)$$

R = nearest neighbour index
n = number of points
$\overline{D}$ = mean nearest neighbour distance
A = area of region or
L = length of line

Note that the **area** (or length) affects the index as well as the mean nearest neighbour distance. Accordingly, defining the area correctly is extremely important. There are two possible conventions for this:

- An area with a well-defined boundary (eg county/administrative district)
  In some cases you may need to carry out an exercise to define the boundary first - eg mapping the CBD.
- The smallest possible rectangle that includes all the points you wish to consider.

If you are intending to **compare** nearest neighbour indices for two regions, you should use the **same** convention for defining the area for each (i.e. both rectangles or both defined boundaries).

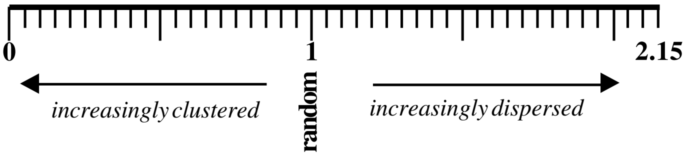
***Exam Hint:-** Make sure you use **comparable units** for your area and nearest neighbour distances - eg m^2 for area and m for distance.*

If the objects you're considering have a noticeable area (eg shops, towns), you also need to consider whereabouts in it you are measuring to. For shops, it should normally be the doorway (since you are interested in the distance people travel to enter the shop), but with towns, you should consider the centre of it.

Interpreting your results

It's generally recommended that you should have at least **30** points before attempting to draw any conclusions from your result; however, if you choose to do a hypothesis test (see below) you can get away with fewer, as the test "allows for" the sample size.

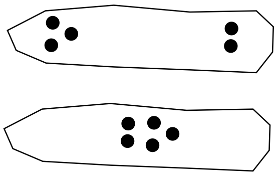
You may find it helpful to plot your result on a scale with 0 at one end and 2.15 at the other. The closer to one end of the line your result is, the more the distribution differs from randomness.



Some points to watch out for:

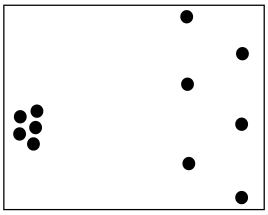
- ♦ If the formula is to be valid, it assumes that points are free to locate anywhere in the region you are considering. This is not always true! For example:
  - distribution of settlements in a region with a large lake
  - distribution of shops around a large tourist attraction such as a castle

- ♦ The formula will not distinguish between regions with one cluster and regions with several. This is particularly apparent if you are using boundaries rather than smallest rectangle to enclose the points.



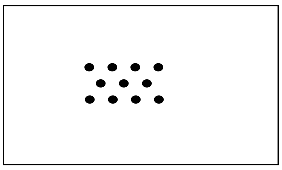
  You may be able to address this by subdividing the region, if there is a legitimate reason for doing so (other than just the distribution of points!)

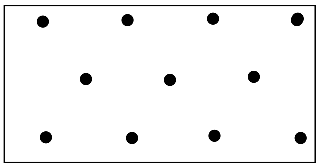
- ♦ A **mixture** of clustered and dispersed patterns may well give a result that appears "random" Again, splitting the region up could be worth considering.



  A combination of separate patterns like this may occur in a town containing different, well-defined shopping regions - for example, one near a tourist attraction and one in the main High street.

- ♦ If you consider the same pattern on different scales you may get very different results. This may be an issue if you are considering an area defined by an administrative boundary or similar, as the official "boundary" may be some way outside the actual town


appears clustered


appears dispersed

Ideas for using Nearest Neighbour Index

- People on beaches
- Visitors at honeypot sites
- Farms in areas with differing geologies
- Trees in a woodland (*planted/natural)*
- Shops/businesses of particular type in CBD
- Distribution of leisure facilities
- Towns/villages
- Distribution of micro-features at a headland
- Clustering of crime/vandalism

Acknowledgements; *This Geo Factsheet was researched and written by Cath Brown.* ***Curriculum Press, Bank House, 105 King Street, Wellington, TF1 1U.*** *Geopress Factsheets may be copied free of charge by teaching staff or students, provided that their school is a registered subscriber.* ***ISSN 1351-5136***

Testing Significance

The significance test involves deciding between your null hypothesis (H_0) and alternative hypothesis (H_1). The null hypothesis here is always:

H_0: The distribution is random

For the alternative hypothesis, you have a choice:

H_1: The distribution is not random —— *non-directional*
H_1: The distribution is clustered
H_1: The distribution is dispersed
directional

The last two versions are referred to as **directional** because they are considering just one specific alternative. These should only be used if you have a sound geographical reason - before collecting any data! - for expecting that alternative. *(For example, if you suspect a certain woodland has been planted, you would be looking for dispersion as your alternative.)* If in doubt, use the first version of H_1.

If you are considering the **non-directional** alternative, you are doing a **2-tailed test**; a **directional** alternative means a **1-tailed test.** This is important when it comes to looking up values in statistical tables.

The **tables** for nearest neighbour have two parts - one for clustered and one for dispersed results.

If your alternative hypothesis is	**You use**
H_1: The distribution is not random	Both tables, 2-tailed test
H_1: The distribution is clustered	Clustered table, 1-tailed test
H_1: The distribution is dispersed	Dispersed table, 1-tailed test

Clustered *Significance levels (0.05 = 5%)* **Dispersed**

1-tailed	**0.1**	**0.05**	**0.025**	**0.01**
2-tailed	**0.2**	**0.1**	**0.05**	**0.02**
2	0.527	0.392	0.276	0.140
3	0.614	0.504	0.409	0.298
4	0.666	0.570	0.488	0.392
5	0.701	0.616	0.542	0.456
6	0.727	0.649	0.582	0.504
7	0.747	0.675	0.613	0.540

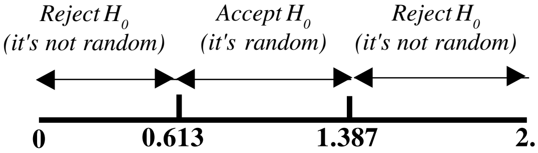
1-tailed	**0.1**	**0.05**	**0.025**	**0.01**
2-tailed	**0.2**	**0.1**	**0.05**	**0.02**
2	1.473	1.608	1.725	1.860
3	1.386	1.497	1.592	1.702
4	1.335	1.430	1.512	1.608
5	1.299	1.385	1.458	1.544
6	1.273	1.351	1.418	1.497
7	1.253	1.325	1.387	1.460

No. of points

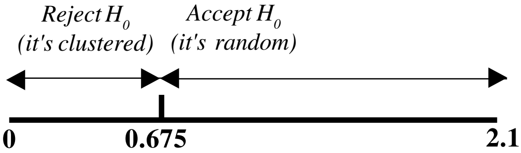
The tables provide a "cut off" value to tell you what counts as "clustered enough" or "dispersed enough" to enable you to reject your null hypothesis. If you don't get a result that is significant, you have to accept your null hypothesis - that the distribution is random.

We will consider doing the test at the 5% significance level with 7 points.

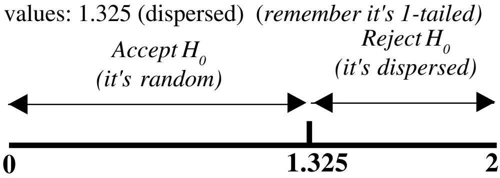
If we used H_1: The distribution is not random:
Tables values: 0.613 (clustered) & 1.387 (dispersed) *(remember it's 2-tailed)*



If we used H_1: The distribution is clustered:
Tables values: 0.675 (clustered) *(remember it's 1-tailed)*



If we used H_1: The distribution is dispersed:
Tables values: 1.325 (dispersed) *(remember it's 1-tailed)*


Accept H_0
(it's random)
Reject H_0
(it's dispersed)
0
1.325
2.15