Detecting Clustering in Point Data Using ArcGIS Pro

Overview of ArcGIS Pro Commands:

The *Analysing Patterns* set of commands within the *Spatial Statistics* toolbox provides 3 options for measuring clustering in both areas and point data. These tools measure clustering of disease, i.e. the tendency for diseased individuals to be clumped together across a whole study area. Whilst they produce an overall measure of clustering across a study area, they do not identify specific locations within the study area where disease rates appear particularly high or low. In contrast, the set of tools within *Mapping Clusters* does enable specific areas with high or low disease rates within a study area to be identified. We will focus here on the *Analysing Patterns* commands.

The *Analysing Patterns* commands are as follows:

Average nearest neighbour: This is a command best suited to point data. It indicates whether a set of points are clustered together, follow a random pattern, or are regularly dispersed across a study area. Averaging these distances for all the points, clustered points will show a low average nearest neighbour distance. Dispersed points that are scattered will show a high average nearest neighbour distance.
Spatial autocorrelation (Moran's I): The (global) Moran's I statistic will indicate whether neighbouring areas have similar disease rates. Although this statistic is most commonly used with polygon data, there are circumstances where you might use this facility with point data.
High/low clustering: This statistic not only indicates whether neighbouring areas have similar disease rates, but it also indicates whether there is a tendency for high disease rates to be grouped together, or low disease rates to be grouped together.
Multi-Distance Spatial Cluster Analysis (Ripley's K Function): We will describe this tool at the end of the exercise.

Data:

The data consist of two leukaemia data files, originally prepared by Prof. Peter Diggle:

(a) **controls** contains the home locations of a group of children who do not have leukaemia.
(b) **cases** contains the home locations of a similar group of children in the same age cohort who have been diagnosed with leukaemia.

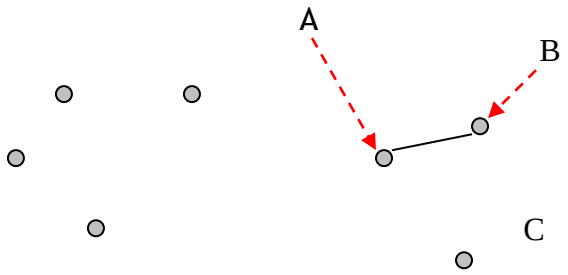
Both data sets are point Shape files and the outline of the study area is also available as a polygon map layer called **leukaemia_studyarea**. The original data are available here: http://www.maths.lancs.ac.uk/~diggle/

Note that the co-ordinates for these data are in 10s of kilometres. As with many other public domain health data sets, the geographic reference system used is not documented.

We also include data relating to John Snow's study of cholera in 19th century London (John Snow's study is sometimes regarded as the first health geography study). This data file is called **snowdeaths** and indicates the places of residence of Londoners who died of cholera.

Using the average nearest neighbour test with point data:

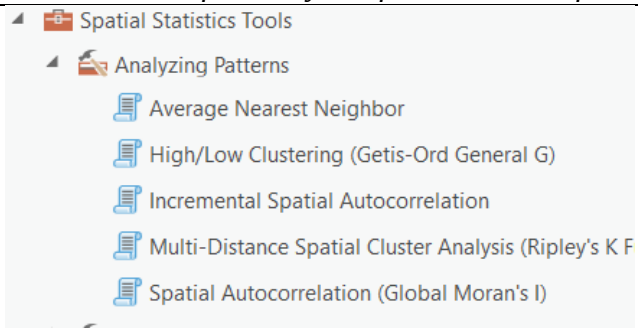
We can use one of these tests – the nearest neighbour test – with some of the point-based health data that we have already seen. This test calculates the distance from each point in a map layer to the nearest neighboring point. For example, in the diagram below, the grey dots represent a point map layer. If we look at Point A, its nearest neighbour is point B, and the distance to this nearest neighbour is shown by the solid line (see diagram below). The software then looks at each point (A, B, C and so on), identifying each point's nearest neighbour and the distance to this neighbour. It then works out an overall average distance, averaging across all the points in the map layer.



In fact, if we assume the points are just randomly scattered across our study area, it is possible to calculate an expected value for this nearest neighbour distance statistically. If the points are clumped together, the average of these distances will be smaller than the value you would expect for a random pattern of points. If the points are evenly spread out across the study area, it will be larger than you would expect for a random point pattern.

Let us see how this works in practice. Load up the cholera deaths data from the John Snow practical as a new map display within ArcGIS Pro. Try:

- From within the *spatial statistics* toolbox, select *analysing patterns* and choose *average nearest neighbour*.



- Choose the Snow deaths as the *input feature class* and for now, leave the distance method set to 'Euclidean Distance' (there is another option here, 'Manhattan distance'. With this option, instead of calculating the straight-line distance between two points, the software calculates distance based on two sides of a right-angled triangle – see diagram below).

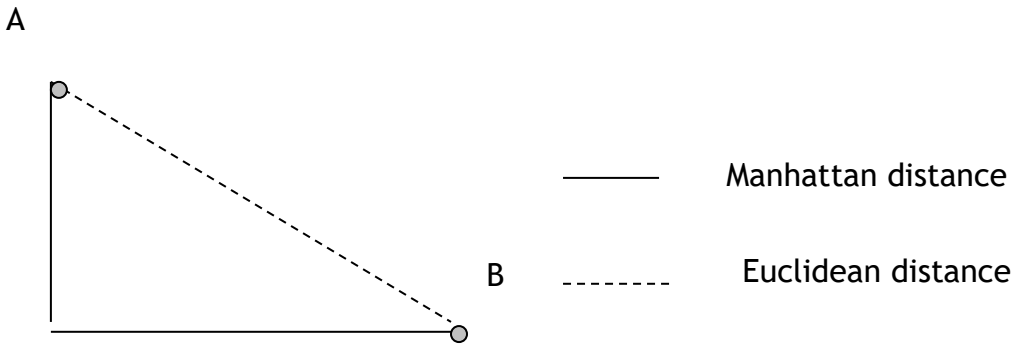
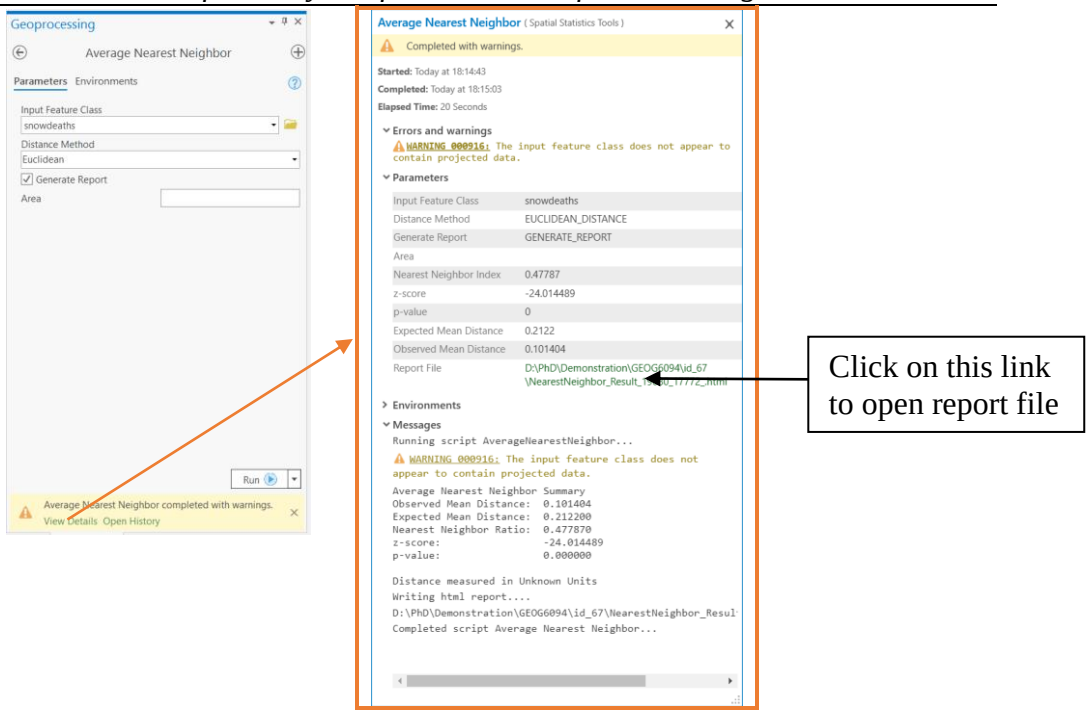


Diagram showing Manhattan and Euclidean (straight-line distances) between two points, A and B).

- Check the box marked '*Generate report (optional)*' and click OK. The tool might take a little time to run, but you should see a message box below *Run* when it has completed – see below. If you click on the words 'View Details', it will bring up the output from the tool (N.B. If you are using ArcGIS 9.x, the output should just display on screen and may look a little different. You may see a green error message, warning you that ArcGIS does not know the map projection for this particular data set, but do not be alarmed!).



What does the output mean?

Underneath the 'average nearest neighbour' section of the output, you will see a number of different figures:

- **Observed Mean Distance**: This is the measured average distance between each death location recorded by John Snow and its nearest neighbouring death. In this case it is 0.10, though as John Snow did not work in 'real' map units, the distance units are not that meaningful here (if instead we had a map layer in State Plane coordinates, this figure would be in metres).
- **Expected Mean Distance**: This is the average distance between each point and its nearest neighbour we would expect, were the death locations spread randomly across the study area, calculated using a statistical formula. It is 0.21 for this data set, so in other words, under random conditions, we would expect each point to be more spread out and further from its neighbour.
- **Nearest Neighbour Ratio**. This is the observed average nearest neighbour distance, divided by the average nearest neighbour distance you would expect if the points were randomly spread. We can read this as follows:
  - If the ratio is round about 1, the observed distance is more or less the same as the one we would expect for a randomly spread point pattern, so this implies the points are falling at random across the study area.
  - If the ratio is less than 1, then the average distance measured from the pattern is less than the average distance we would expect with a random spread of points. Our points are generally much closer together than we would expect if they were spread randomly, suggesting they are clustered.
  - If the ratio is more than one, then our point pattern is more

evenly spread across a study area than we would expect – the average distance we see in our pattern is bigger than that we would expect from a random spread of points.

In this example, the ratio is 0.477 (=0.10 / 0.21), meaning the average nearest neighbour distance we see for the John Snow deaths is around half what we would expect, had the same points been scattered randomly across central London. This looks like a clustered point pattern.

Two more figures tell us whether our points are so close together that this could not have happened by chance – what we would call statistical significance :

- **Z-Score** is something called the Z Score for the nearest neighbour ratio. This is the most difficult concept to explain within the test, but in essence, one random pattern will likely look different from another random pattern (that is what randomness is after all!). Because random patterns can come out differently from one another, the average nearest neighbour distance for a random pattern is not always exactly the expected value. Sometimes it could be a little higher, sometimes a little lower. The Z score statistic tries to measure this random spread we get around the expected nearest neighbour distance value[1]. For a clustered pattern, the Z score will be negative, for an evenly spread pattern, it will be positive, and for a random pattern it will be close to zero. Negative Z scores less than around 2 suggest that our nearest neighbour distance is significant and very unlikely to have come from a random pattern. Positive Z scores greater than around 2 suggest that points are much more evenly spread that we would expect by chance.

In this case, our **Z-Score** is **-24**, which means that it is really, really unlikely to have come from a random pattern.

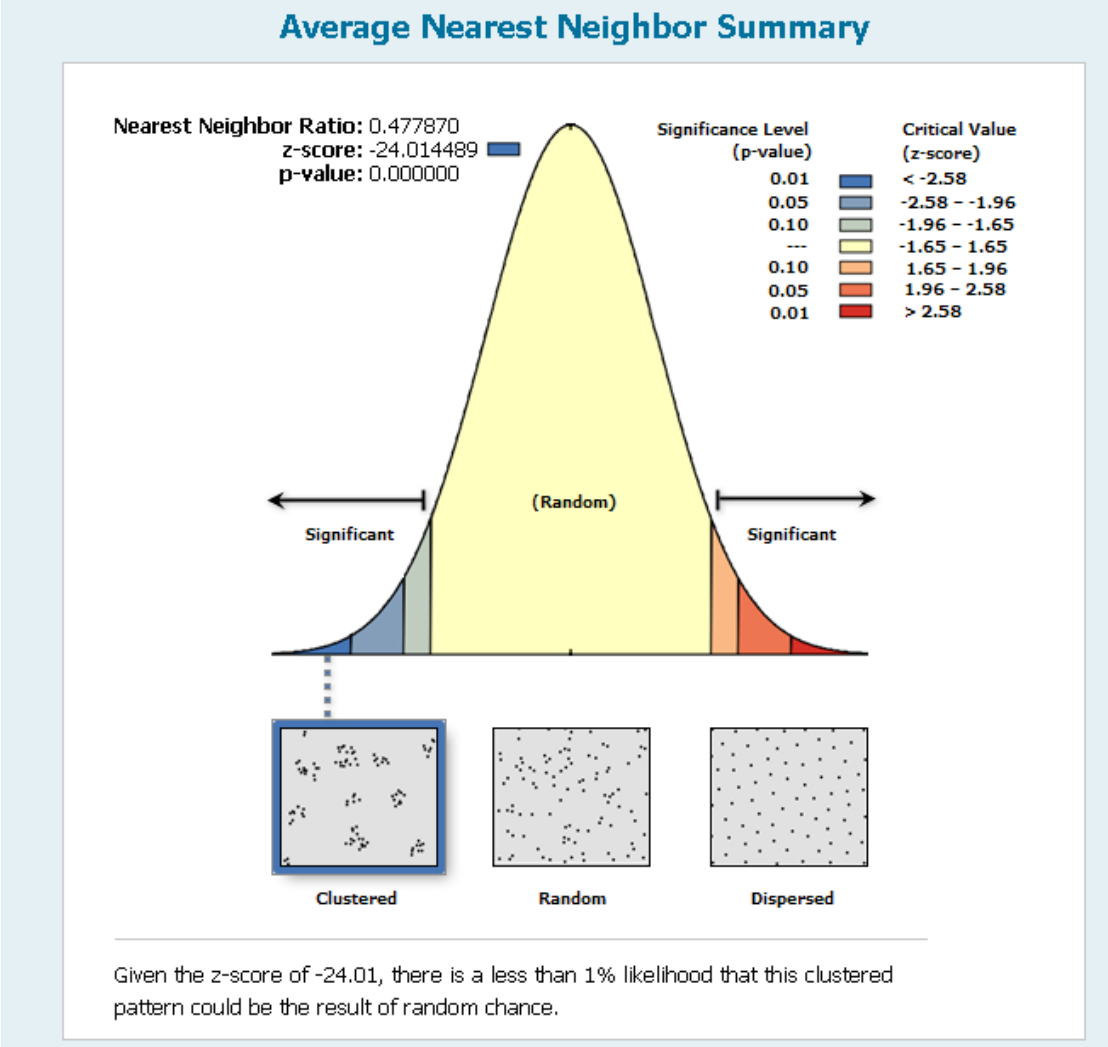
- **PValue** is the probability of seeing a more clustered pattern than we have observed by chance alone. Because it is a probability, it will be between zero and one. Numbers very close to zero (e.g. <0.025) suggest that the observed clustering is very unlikely to have happened by chance – what we would call significant clustering. Numbers very close to one (e.g. >0.975) suggest that we would nearly always expect a random pattern to be more clumpy than the one we are dealing with – what we would call significant dispersion. Numbers somewhere in the middle (e.g. 0.3; 0.5, 0.7 etc) suggest that whilst our average nearest neighbour distance is higher or lower than what we would expect for a random pattern, this could quite easily have happened randomly – in other words there is no significant clustering or dispersion.

Our **P value** for the John Snow deaths is **0** so there is really no chance that such a clumped point pattern could have arisen by chance (though see the caveats at the end of this handout).

If you click on the **html report file (see image above)** in the output, this same information will be displayed in a graphical format in your browser,

[1] Technically, it tells us how far the observed nearest neighbour distance is from the expected one, measured in terms of the standard error of the expected value.

as shown below.



You should see a dialog box appear, showing the results of the nearest neighbour statistic analysis. When you have viewed the results, click on OK to close down this graphical output box.

Task 1: Can you think of any reasons why you might want to be cautious in interpreting the nearest neighbour statistic for John Snow's data? (see the end of this worksheet for some ideas)

Now remove the cholera deaths data set. Try loading up the leukaemia case and control data into ArcGIS. Notice that with this data set, we do know something about the geography of the underlying population at risk. The control data are free of disease and therefore give us an indication of the spatial spread of the population at risk.

Try running the average nearest neighbour statistic for the cases.

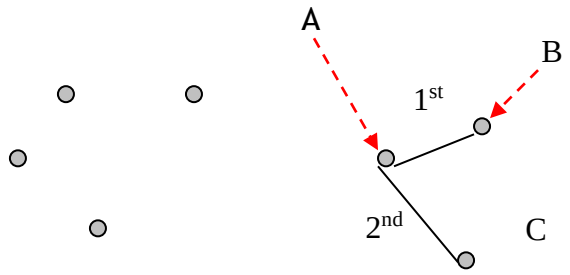
Make a note of the value of the nearest neighbour statistic for the disease cases: _____

Try running the average nearest neighbour statistic for the controls. Make a note of the value of the nearest neighbour statistic for the disease-free controls: _____

Task 2: Is there any evidence that the disease cases are more clustered together than the disease-free controls?

Summary and Further Ideas:

Not all cluster statistics work exactly like the nearest neighbour statistic. However, they will have two features in common. In general, you will often see a test statistic (like the Nearest Neighbour Ratio here), a summary number that is calculated to describe a pattern either within a portion of a study area (local) or across an entire study area (global). You will also see a P value (just like the one we saw here), that shows how likely it is that this test statistic value could have come about by chance.



We mentioned the **multi-distance spatial cluster analysis** tool at the start of this exercise, which sits in the same part of the geoprocessing panel under *analysing patterns* (See first image on page 2). In theory, you could imagine going further with nearest neighbour analysis. Instead of simply considering the nearest neighbour to each point, you could imagine by-passing the nearest neighbour and visiting the next-but-one nearest neighbour, then calculating the distance to that point. For point A above, the next-but-one neighbour might be C and we could work out a distance to this second neighbour as well as the first, nearest neighbour. We could then run through exactly the same calculation as we just did, but figure out an average next-but-one-nearest neighbour distance. We could keep going and look at the third nearest, fourth nearest, and so on.

What this would do is tell us not only whether points were clustered,

but also whether the points were clustered into one large cluster or many small clusters – so how big the clusters are. It is this type of analysis that can be undertaken with the more complex **multi-distance spatial cluster analysis** tool.

Answers / ideas for Task 1:

Some potential problems with the John Snow data set in this analysis are:

- It is unclear how far the original investigation of deaths was itself clustered in terms of effort on the ground. Could the clustering simply reflect the geographic concentration of effort in mapping locations of deaths within Soho?
- Notice that we do not know whether any clustering of cholera deaths is because the underlying population is clustered, or whether the method of recording the deaths may have led to clustering in this data set.