Practical exercise: point pattern analysis of contamination sites

The Known Contaminated Sites List (KCSNJ) for the state of New Jersey in the USA represents non-homeowner sites and properties within the state where contamination of soil or groundwater has been confirmed at levels equal to or greater than applicable standards.

Data for this exercise were retrieved from the State of New Jersey Department of Environmental Protection (http://www.state.nj.us/dep/gis/stateshp.html). The data file consists of a time-series of known contaminated soil or groundwater sites for 2004-2013 in New Jersey. Attribute data consist of location information, date when incident recorded and the category of site contamination: A = sites with onsite sources of contamination; B = sites with unknown sources of contamination; C = sites closed with restrictions.

Can you answer the following questions with the data provided:

A. Are the contaminated sites data clustered?
B. Is there any difference between the amounts of clustering for each category of site contamination?
C. Is there a difference in the spatial pattern of sites identified before and after 2013?
D. Is there any difference between data patterns for Burlington county versus Hudson county?

Implementing a nearest neighbour analysis

1. Open ArcGIS and add in the contaminated sites dataset.
2. Navigate to the spatial statistics toolbox > analysing patterns > average nearest neighbour. Open the tool.
3. Select *ContaminatedSitesNJ* as the input file. Tick the *generate report* box to produce a graphical summary. Leave the options as default and click ok.
4. Results of nearest neighbour analyses can be found in the *Results* window. Navigate to this in the *Table of Contents* window. If you are unsure how to find this, click on the *geoprocessing* menu and then click on *results*.
5. Double-click on the *report file* which will open in your web browser. This provides a nice graphical summary of the analysis as well as the output statistics.

Tip: If the *average nearest neighbour* tool generates an error message, you may need to disable background processing for geoprocessing. To do this, head for the *geoprocessing* menu, select *geoprocessing options*, and uncheck *enable* under *background processing.*

What does the output mean?

The Average Nearest Neighbour tool returns five values: observed mean distance, expected mean distance, nearest neighbour index, z-score, and p-value. The observed mean distance is the average distance between each point and its nearest neighbour in your point pattern. The expected mean distance is the average distance between neighbouring points that you would expect, were the

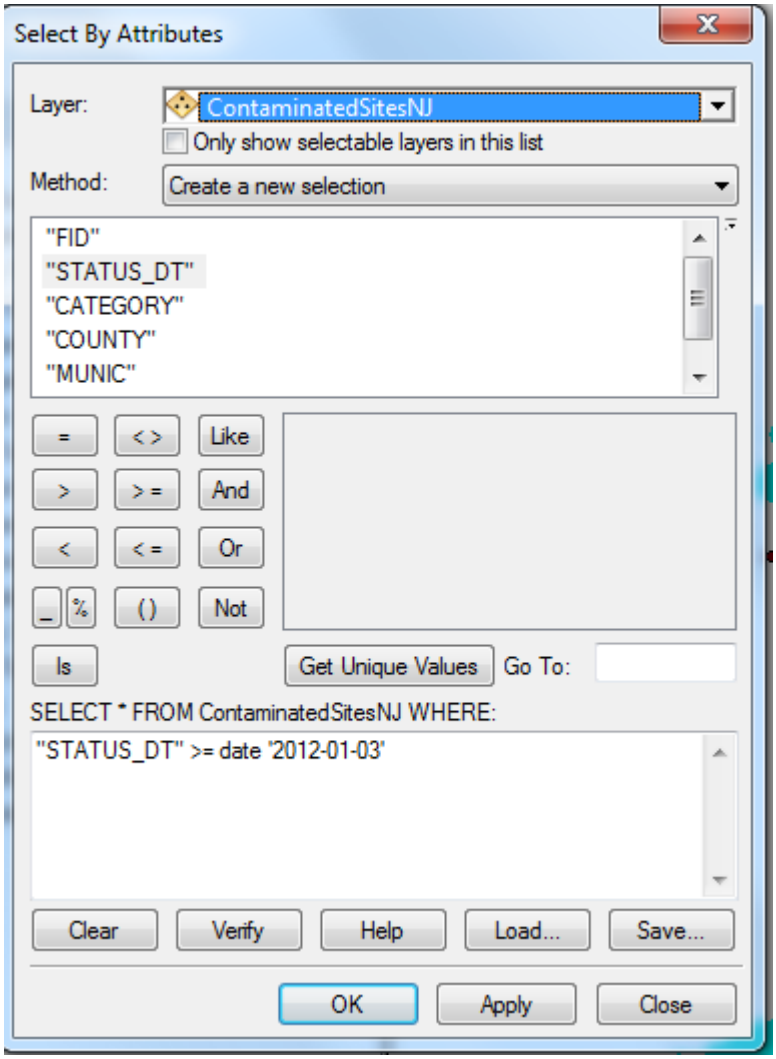
pattern to be random. The nearest neighbour index is simply the ratio of these two numbers, i.e. the observed nearest neighbour distance divided by the expected distance. If the index (average nearest neighbour ratio) is less than 1, the pattern exhibits clustering because the observed distance between neighbouring points is smaller than would be expected under randomness. If the index is greater than 1, the trend is toward dispersion, because points are more widely spread than would be expected under randomness. The P-value indicates the probability that the pattern is random. A value less than 0.05 would suggest that the pattern is clustered, whilst a value greater than 0.95 would suggest that the point pattern is uniformly distributed. P values between these extremes suggest a randomly distributed point pattern.

The Z-score compares the observed nearest neighbour distance with the nearest neighbour distance expected for an equivalent random pattern. A negative Z-score indicates that the observed average nearest neighbour distance is lower than the expected distance and is suggestive of clustering, whilst a positive Z-score indicates the opposite. A Z-score of zero indicates that the observed average nearest neighbour distance is the same as would be expected under randomness. For any given random pattern, the average nearest neighbour distance may not match the expected value under randomness. Sometimes it may be higher or sometimes lower than the expected value under randomness. This spread in average nearest neighbour distance values under randomness can be measured in terms of standard deviations. The magnitude of the Z-score indicates how far the observed average nearest neighbour statistic lies away from its expected value in terms of standard deviations. Thus, a Z score of -1 means the observed average nearest neighbour statistic is one standard deviation lower than its expected value.

Are the contaminated data clustered, random or dispersed?

Generating average nearest neighbour statistics for subsets of data

Now have a go at filtering the attribute data in the *ContaminatedSitesNJ* file to answer questions B to D above. For example, to filter the subset of sites that have been identified since the beginning of 2013 can be identified by heading for the *selection* menu and choosing *select by attributes*, as shown below:



Having filtered the dataset to pick out the sites that have been identified only recently, these can either now be processed using the *average nearest neighbour* tool, or the selected sites can be saved permanently as a new map layer by right-clicking on the layer and selecting *data* and then *export data*. The newly saved map layer can then be used as an input to the nearest neighbour tool.

What do you think may be the underlying causes of contamination and associated patterns in the data? What vital attribute data do you suppose you are missing to make inferences about the level of contamination?

Additional information

Further details on parameters required for using the nearest neighbour tool can be found here.
If you are keen to explore further functionality and application of analysing patterns using the spatial statistics tool, then an ESRI example application tutorial for mapping spatial patterns of dengue fever can be found here.