Geographical Cluster Hunter

Summary

Professor Stan Openshaw developed the Geographical Cluster Hunter. It has been used in a variety of geographical search problems such as to analyse child leukaemia cases near a nuclear reactor (Openshaw *et al*, 1988). The algorithm is a sequential search engine. It looks for concentrations of events within a specified distance of a series of search points. The distance radii of the search area vary between a minimum and maximum value in sequential steps, all parameters defined by the user. The input for the algorithm requires a base population, for example all of the geocoded tweets posted over the opening ceremony of the Olympics. In addition the algorithm also requires the identification of 'events' of interest, for example geocoded tweets posted over the opening ceremony of the Olympics that mention the word 'medal'. In this case the events set is a subset of the base population but this does not always have to be the case. The algorithm would work in a scenario where a population of individuals made shopping trips. Some individuals may make more than one shopping trip and, therefore, shopping events could outnumber the number of individuals in the base population.

Sequential search

The algorithm starts its search in the northwest corner of the search area, determined from the extent of the input dataset. It begins with the smallest specified search radii and progresses along the search area to the east allowing each search circle to overlap by the user specified amount. When the western extent is reached the search moves back to the west and to the south by one overlap increment, as shown in Figure 1 below.

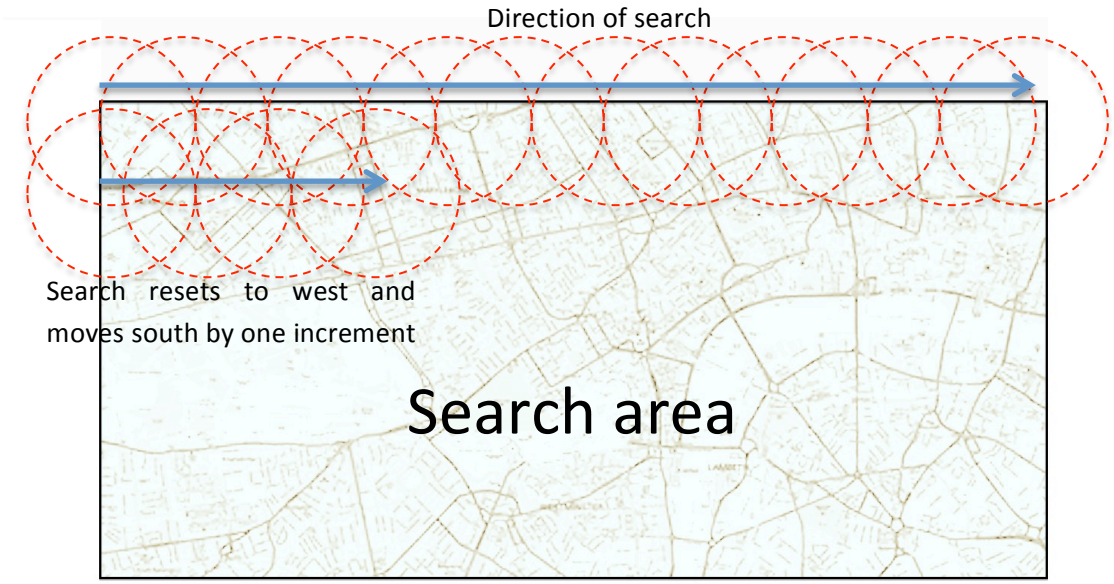

Figure 1: search coverage by the cluster hunter algorithm

Once the whole search area has been covered the size of search circle is increased by the user-defined step and the search begins again. This process is repeated until the search has been completed using the maximum search radius specified by the user.

The number of observed and expected events within the search circle is compared. Expected events are calculated as an indirectly standardised estimate based on the underlying background population if automatic standardisation is selected, see equation 1.

Equation 1:

$$expected\ events\ = \frac{\sum events}{\sum base\ population} \times base\ population\ within\ search\ circle$$

If more cases are found than are expected a Poisson test is performed to see if the difference is statistically significant against a user defined threshold. If the test is significant the search circle is saved as a potential cluster area. All of the event identifiers from the contributing events are also saved with the search circle details for subsequent analysis.

Parameter settings and results

Inputs

The inputs required are a unique record identifier, x and y coordinates, event counts and population counts at the relevant x, y coordinate. The table containing the data should be dragged to the input table text area and then the required fields selected from the five drop down boxes below. All fields should be numeric (imported as either Double or Integer). Fields imported as text will not be displayed in the field selection drop down boxes.

For detailed instructions on linking data into the Flexible Modelling Framework (FMF) see http://eprints.ncrm.ac.uk/3177/2/microsimulation_model.pdf, sections 2 – 5 deal with installation and general use of the FMF.

The algorithm will process pre-aggregated data or it will aggregate individual data. Therefore, it is valid to enter input data with several individual population or events occurring at the same geographical location, these will be aggregated prior to the algorithm calculations commencing. Furthermore, it is valid to have events occurring at a location without a corresponding population count, this enables the researcher to consider events in the context of but not semantically linked to the background population of interest. It is also valid to consider events and populations recorded at different geographical resolutions so long as they are on a consistent accepted coordinate system.

Outputs

The outputs from the search will be saved in a table in the specified data source. To specify the data source drag and drop the required data source folder from the data sources tab to the data source text area in the 'Output Table' section of the dialog. The table name will default to 'significant clusters'; the user can change this. If the table name selected already exists a number will be suffixed to the end of the table name. For example if 'significant clusters' already existed in the data source the table name would become 'significant clusters_1', if this already existed it would become

'significant clusters_2' etc. The output from the algorithm consists of five fields for each search circle that returned a significant difference when tested using the Poisson significance test (discussed below):

1. ID – the unique identifier of the cluster.
2. X – the x coordinate at the centre of the search circle.
3. Y – the y coordinate at the centre of the search circle.
4. Radius – the radius of the search circle that returned this result in metres.
5. Value – the difference between the observed and expected value for the search circle.

If no significant results are found the table will be returned empty. The algorithm output can readily be imported into a desktop GIS system such as QGIS or MapInfo for display or further analysis.

Coordinate Type

It is possible to set the coordinate type for the input x, y geographical coordinates. Currently, WGS84 latitude and longitude and British National Grid coordinate types are supported. Please ensure that you select the correct coordinate type an incorrect selection will result in spurious results and potentially lengthily algorithm run times.

Search circle

The ability to set a range of values for search circle radii allows the user to sweep a range of resolutions searching for clusters. The user needs to consider the resolution of the search being carried out and the scarcity of the data over the geographical area when adjusting these settings. With sparse datasets or large geographical areas, the minimum and maximum radii for the search circle may need to be increased from the default settings. However, with local small geographical scale datasets they may need to be decreased from the default settings:

- Minimum search radius = 500 metres
- Maximum search radius = 5,000 metres
- Search radius increment = 500 metres

Statistic

The minimum number of events and population settings can be adjusted. These settings control the lowest level that a significance test will be performed on a search circle. If the count of events or population in the search circle fall below the specified corresponding value a significance test will not be performed and the search circle will not be saved. The default value for both of these parameters is 1.

The significance threshold for the Poisson test is set to 0.0099 by default. The higher the value the more relaxed the significance test, the lower the value the more stringent the significance test. The significance test finds the probability that the observed events are the same as that the expected events according to the Poisson distribution. If the probability is lower than the set threshold the 'null' hypothesis (the observed and the expected are not significantly different) is rejected. This means that the difference is statistically significant at the specified threshold under the Poisson distribution.

The ‘standardise’ checkbox enables the user to disable the automatic standardisation process. If this checkbox is left checked (the default) the algorithm will apply the calculation in equation 1 to the results for each search circle to estimate the expected number of events. However, if a more sophisticated standardisation process has been applied to the data then the default calculation is no longer required and the checkbox should be unchecked. Any pre-standardised expected values should be entered into the cluster hunter algorithm in the population field is automatic standardisation is not required.

Interpreting results

Interpreting the results should be undertaken with care. The first point to make is that the indirect standardisation undertaken within the algorithm does not control for any attributes within the base population, such as age, ethnicity or gender etc., to account for these the user needs to pre-standardise the values and deselect the standardise parameter checkbox. The design and any pre-processing of the input base population and events needs to be thought through to ensure that the results produced are sensible in the context of the analysis.

In general, the higher the difference between the expected and observed cases with relation to the size of the search circle, the smaller the circle, the denser the potential cluster the more important the cluster may prove to be. Again, caution must be exercised when considering the size of the search circle against the resolution of the underlying data. Setting the minimum search circle size too low may result in many small spatially distributed clusters that appear significant. The ability to set a range of search circle radii allows the researcher to consider potential clusters at a number of spatial resolutions in one pass of the data. A cluster that presents at several different spatial resolutions is a more robust option for further investigation than one that is present at one resolution only.

References

Openshaw S, Charlton M and Craft A (1988) Searching for leukaemia clusters using a Geographical Analysis Machine. *Regional Science Association* 64: 95-106