Automated Generation of Hypothesis of Processes Causing Clusters

Jamison Conley, PhD

Department of Geology & Geography, West Virginia University

OBJECTIVE

There are plenty of computational and statistical methods for detecting spatial clusters, although the interpretation of these clusters is a task left to the user. This research develops computational methods to not just detect, but also analyze the cluster to hypothesize one or more potential causes.

BACKGROUND

Computational and statistical methods for detecting disease clusters, such as the spatial scan statistic [1], have become frequently used tools in epidemiology. However, they simply tell the user where a cluster is, and leave the analysis task to the user. Multivariate visualization tools provide one way for this analysis, e.g. [2]. The approach developed in this research is computational in nature, using computer vision techniques to analyze the shape of the cluster. Shapes are used here because different spatial processes that cause clusters, such as pollution along a river, create clusters with different shapes. Thus, it may be possible to categorize clusters by their respective spatial processes by analyzing the cluster shapes.

METHODS

Figure 1 diagrams the components of the software system developed in this research. It starts with a cluster detection module which simply detects the clusters. In the tests of this system, GAM/K [3] is used because it generates a cluster surface, not just a collection of circles. The shape extraction module finds the shape of the cluster by taking a threshold of the surface given by GAM/K. For continued computational analysis, the shape from this threshold is converted into a numerical format for statistical representation. This research uses a set of seven image moment invariants [4] as that representation. This statistical representation is then compared against a database of cluster shapes using the pattern matching module.

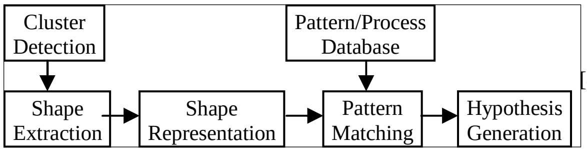

Figure 1 – Diagram of the software system.

The pattern/process database contains these shapes, each of which is stored in the same statistical representation, and each of which is labeled with an associated spatial process. For example, a long, narrow, winding shape associated with a river would be labeled with the process of disease vector dispersion within a river. The pattern matching module compares the shape of the detected cluster with the shapes in the database to find the closest matches. This research uses the k-nearest neighbor algorithm.

The hypothesis generation module uses Dempster-Shafer evidence theory [5] to turn the results of the pattern matching into a hypothesis like "this cluster is most likely caused by pollution in a river." This works by taking the results of the pattern matching for several shapes derived from a single cluster, such as a series of shapes from multiple thresholds of the cluster surface from GAM/K, and combining the pattern matching results into a single cluster hypothesis using combination operators from evidence theory.

RESULTS

This system was tested against shapes derived from GIS files of rivers, roads, urban areas, and soil regions. The system's ability to distinguish between these classes of shapes, while far from perfect with an accuracy rate between 55% to 60% correct depending on the *k* value used in the k-nearest neighbor algorithm, is considerably better than random.

CONCLUSIONS

The ability of the software system developed in this research to distinguish between different geographic shapes demonstrates the potential of shape analysis to assist epidemiological investigations by hypothesizing potential causes of clusters.

REFERENCES

[1] M. Kulldorff, "A Spatial Scan Statistic," *Communications in Statistics, Theory and Methods,* vol. 26, pp. 1481-1496, 1997.

[2] F. Hardisty, "The GeoViz Toolkit," in *Auto-Carto,* Las Vegas, NV, 2005.

[3] S. Openshaw and A. Craft, "Using the Geographical Analysis Machine to Search for Evidence of Clusters and Clustering in Childhood Leukaemia and Non-Hodgkin Lymphomas in Britain," in *The Geographical Epidemiology of Childhood Leukaemia and Non-Hodgkin Lymphoma in Great Britain 1966-1983,* ed. G. Draper, pp. 109-122, London, HMSO, 1991.

[4] M.-K. Hu, "Visual Pattern Recognition by Moment Invariants," *IEEE Transactions on Information Theory*, vol. 8, pp. 179-187, 1962.

[5] G. Shafer, *A Mathematical Theory of Evidence,* Princeton, NJ: Princeton University Press, 1976.

Further Information:

This work was done while the author was at Pennsylvania State University.
Jamison Conley, Jamison.Conley@mail.wvu.edu
Source code available at http://code.google.com/p/clusterer/