

## A Genetic Approach to Automated Cluster Detection in Point Datasets

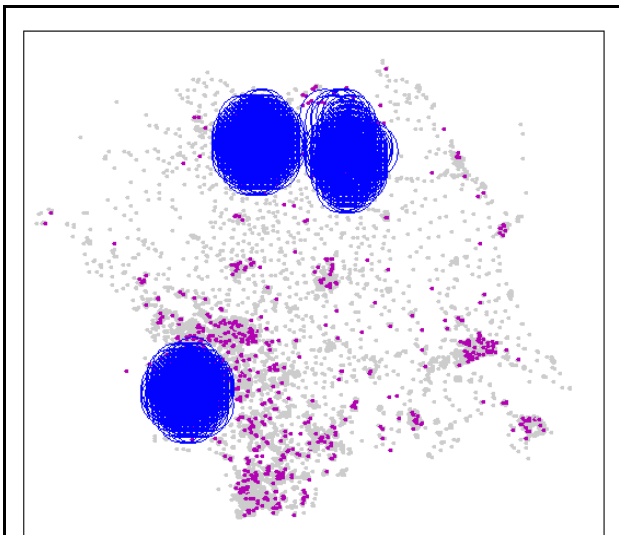
**Jamison F. Conley**

GeoVISTA Center, Department of Geography,  
The Pennsylvania State University  
302 Walker Building  
The Pennsylvania State University  
University Park, PA, 16802, United States of America  
jfc173@psu.edu

The amount of geographic data has rapidly increased throughout the past decade. Impressive data collection capabilities combine with the ability to merge datasets, resulting in a rapid growth in both the number and size of geographic datasets. Meanwhile, the number and urgency of social, environmental, and public safety applications that use these datasets is so great that it has been argued that “there is an undisputed national imperative for analysis” of these datasets (Openshaw 1995, p. 4). One way to analyze these datasets is by detecting clusters. However, many cluster analysis methods do not scale well to large datasets. In addition, some of these methods return many nearly-identical results for each cluster, which unnecessarily adds to the post-processing burden. Therefore, it is necessary to develop new methods and algorithms to perform cluster analysis on large datasets.

Many algorithms for automated cluster detection have been proposed within geography (Openshaw *et al.* 1987, Besag and Newell 1991, Fotheringham and Zhan 1996, Openshaw and Perrée 1996) and in other disciplines (Kaufman and Rousseeuw 1990, Ester *et al.* 1996, Zhang *et al.* 1996). However, each of these algorithms has its faults; some

ignore the spatial component of data that is important in geography (e.g. PAM and CLARA, Kaufman and Rousseeuw 1990), others assume the researcher has hypotheses about the number, size, and/or location of clusters (e.g. Diggle 1990), and still others do not scale well to large datasets (e.g. GAM, Openshaw *et al.* 1987). Finally, some methods (e.g. GAM) return so many results for potential clusters that it is challenging for the user to find the one that best represents each cluster. This is shown in Figure 1, where GAM averages 200 circles per cluster, making it difficult to find the best description of each cluster. It is preferred that a cluster detection method returns only one representation

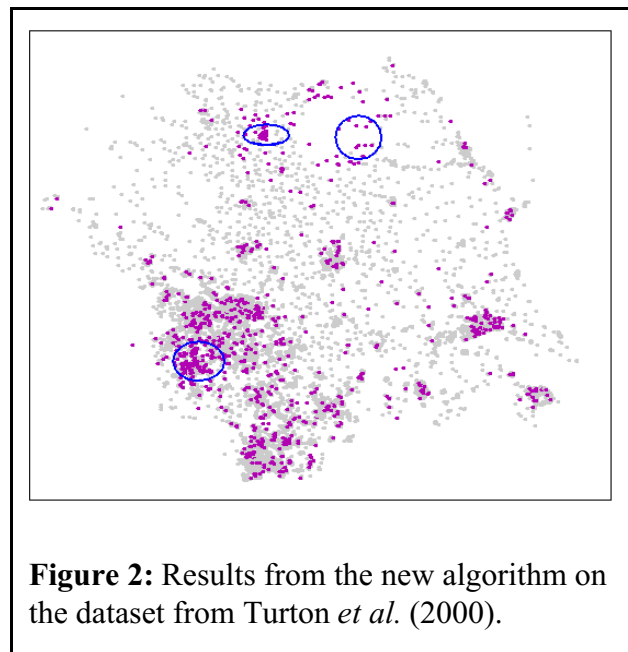


**Figure 1:** Results for GAM on the dataset from Turton *et al.* (2000).

of each cluster. It is highly unlikely, if not impossible, that any single cluster detection algorithm could completely overcome every one of these obstacles. Even so, new algorithms can still improve upon the performance of previous algorithms.

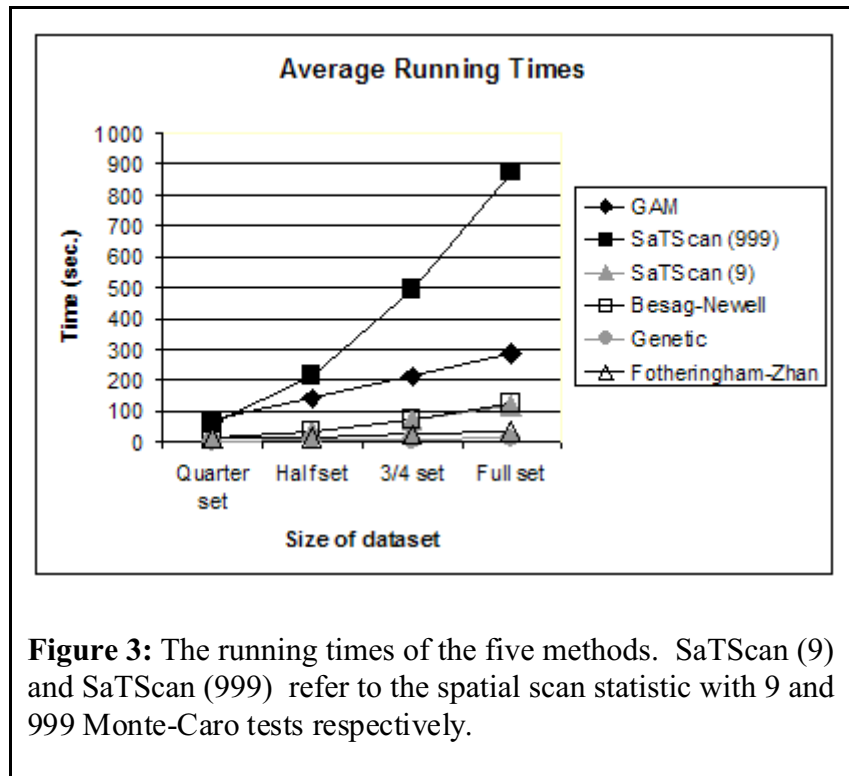
In this poster, I present a new algorithm for cluster analysis. The algorithm should meet the following criteria: (1) it accounts for the spatial nature of geographic data; (2) it contains as few and as general *a priori* assumptions as possible; (3) it scales to large datasets better than existing algorithms that meet criteria (1) and (2); and (4) it returns one representation of each cluster.

To achieve these goals, I develop an algorithm that is based on a genetic algorithm (GA) and GAM (Openshaw *et al.* 1987). GAM is the starting point because it already meets the first two criteria. The primary assumptions that GAM makes are that the clusters are approximately circular, and that the statistical definition of a cluster that it uses is valid. Later algorithms based on GAM (Besag and Newell 1991, Fotheringham and Zhan 1996) attempt to improve both the time complexity of GAM and the quality of the results. Also, there has been a previous attempt to develop a GA based on GAM: MAPEX (Openshaw and Perrée 1996). However, the GA behind MAPEX tended to converge on a single cluster even if there were many clusters in the dataset, making it unable to detect more than one cluster at a time. The algorithm presented here overcomes this convergence. Also, while GAM and the other methods assume that the clusters can be represented by circles, the method presented here represents clusters as ellipses, which reduces the shape-based bias, although it does not entirely eliminate it. Finally, while GAM tends to return a very large number of circles per cluster, the new algorithm only returns one or two ellipses per cluster, as shown in Figure 2.



**Figure 2:** Results from the new algorithm on the dataset from Turton *et al.* (2000).

I compare the new algorithm with four existing algorithms: GAM, case-point centered searching (Besag and Newell 1991), random searching (Fotheringham and Zhan 1996), and the spatial scan statistic (Kulldorff 1997) by applying them to a point dataset of simulated disease incidence developed by Turton *et al.* (2000). The algorithms find regions of the dataset where the rate of disease incidence is higher than the average rate.



Comparisons between the proposed method and the other four algorithms show that the new algorithm is as capable of finding clusters as the other methods, although it is less consistent due to the stochastic nature of a GA's searching mechanism. Also, as shown in Figure 3, its running times are faster than the other algorithms. Unlike MAPEX, the proposed algorithm does not converge on a single cluster, but is able to find most, if not all clusters in the dataset. In Figure 2, all three clusters present in the dataset were detected by the algorithm. Finally, because it searches for elliptical clusters in addition to circular clusters, it can capture a wider variety of geographic clusters.

The algorithm proposed here meets all four goals: it successfully accounts for the spatial nature of geographic data, it can be operated with minimal prior knowledge about the dataset, it scales to large datasets, and it returns fewer ellipses per cluster.

### Acknowledgments

I would like to thank Mark Gahegan and James Macgill for much help in developing this research. This research was funded in part by the National Cancer Institute grant 54K3.

### References

Besag, J. and J. Newell (1991). "The detection of clusters in rare diseases." *Journal of the Royal Statistical Society, Series A*. 154:143-155.

- Diggle, P. J. (1990). "A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point." *Journal of the Royal Statistical Society, Series A*. 153:349-362.
- Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proceedings of the Second International Conference on Data Mining KDD-96*. Portland, OR. 226-231.
- Fotheringham, A. S., and F. B. Zhan (1996). "A comparison of three exploratory methods for cluster detection in spatial point patterns." *Geographical Analysis*. 28:200-218.
- Kaufman, L., and P. J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc.
- Kulldorff, M. (1997). "A spatial scan statistic." *Communications in Statistics: Theory and Methods*. 26:1481-1496.
- Openshaw, S. (1995). "Developing automated and smart spatial pattern exploration tools for geographical information systems applications." *Statistician*. 44:3-16.
- Openshaw, S., M. Charlton, C. Wymer, and A. Craft (1987). "A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets." *International Journal of Geographic Information Systems*. 1:335-358.
- Openshaw, S. and T. Perrée (1996). "User-centred intelligent spatial analysis of point data." in Parker, D., ed., *Innovations in GIS 3*. London: Taylor & Francis. pp. 119-134.
- Turton, I., S. Openshaw, C. Brunsdon, A. Turner, and J. Macgill (2000). "Testing space-time and more complex hyperspace geographical analysis tools." In Atkinson, P. and D. Martin, eds. *Innovations in GIS 7*. London: Taylor & Francis. pp. 87-100.
- Zhang, T., R. Ramakrishnan, and M. Livny (1996). "BIRCH: An efficient data clustering method for very large databases." in *Proceedings of SIGMOD '96*. Montreal, Canada. 103-114.