# Putting the Geographical Analysis Machine on the Internet Revisited

Ian Turton[1] and Andy Turner[2]

[1]GeoVISTA Center, Department of Geography, The Pennsylvania State University, University Park, PA 16801 ijt1@psu.edu
[2]The Centre for Computational Geography, School of Geography, University of Leeds, Leeds LS2 9JT A.G.D.Turner@leeds.ac.uk

## Introduction

Openshaw et al. (1999) wrote about their experiences putting the Geographical Analysis Machine (GAM) on to the Internet to allow a wider use. Their motivation was that, at the time, proprietary GIS programs lacked sophisticated geographical analysis technology. The experimental system that was developed relied on a complex set of Unix scripts and FORTRAN code and proved to be unsustainable. The system also required a text file containing the locations of population and case points which was difficult to create for a naïve user. Lastly it required users to upload these potentially confidential information to a remote server trusting the server administrators who were unknown (but obviously trustworthy). In the last decade, since the original attempt to put GAM on the Internet, the situation appears not to have improved. There are still demand for more user friendly and secure ways of exploring geographical data for evidence of spatial clustering. Robertson and Nelson (2010) state that "...training and software availability were cited as the primary barriers to the uptake of space-time disease surveillance..." and provide a general assessment that for the programs they tested - handling the data formatting was difficult and the interpretation of outputs was challenging.

## Motivation

This paper seeks to solve the same problem that Openshaw et al. (1999) attempted, making use of modern developments in cloud and grid computing, distributed spatial data management and improved computing power. From the literature (*e.g.* Olsen et al., 1996; Robertson and Nelson, 2010) there is a demand from epidemiologists for a simple system that will: import their case data; import population data (preferably from a Census site directly, or from files they download from one); and, exports a geographically referenced, easy to understand map of the potential clusters for them to investigate.

As with anyone handling confidential data, epidemiologists are concerned about data security. Any system that is to be used with confidential data needs some form of guarantee that the data will be secure and will not become available to others (at least not provided without clear usage restriction and only to other users of those confidential data). To guarantee the security of a software system running on a networked machine, the software source code needs to be inspected

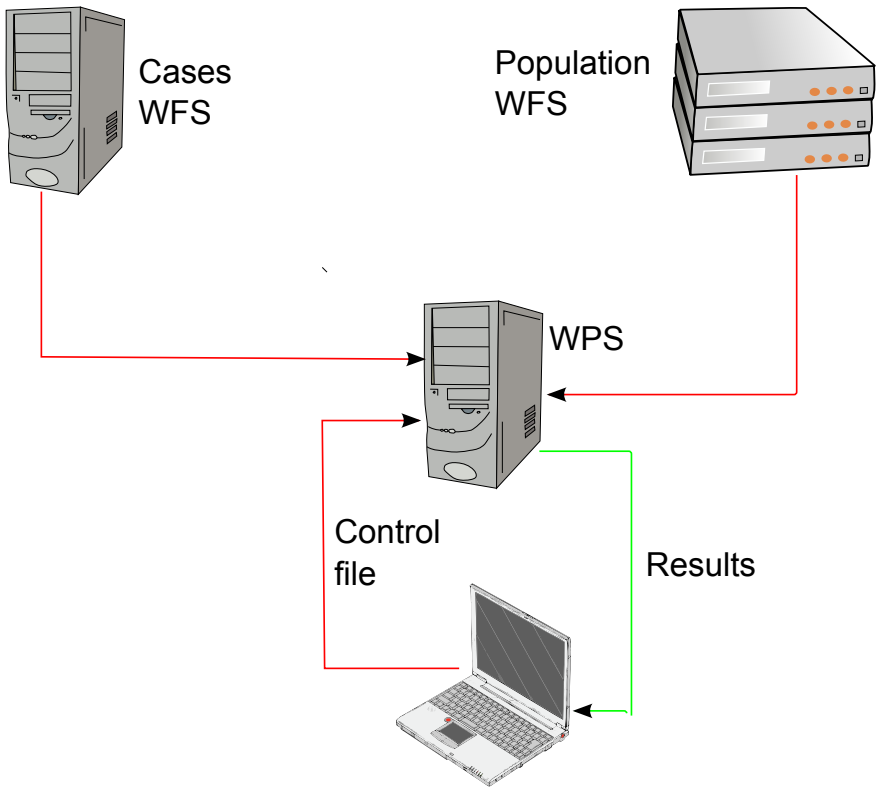


Figure 1: Implementation Diagram

and the authority for a guarantee trusted. A system that is completely open source and based on standards compliant open source service components allows anyone to inspect the source code and this helps to verify that data is secured in the system. Additionally, being open source allows for the academic rigour of the algorithm implementations to be assessed. In a closed source system, arguably there is too great a risk with respect to confidential data security.

## Implementation

The system described is programmed in Java using the GeoTools library (Turton, 2008). The system makes use of the Web Processing Standard (WPS) (OGC, 2007) as implemented by GeoServer (an open source server which implements other important Open Geospatial Consortium (OGC) standards). Users are able to seamlessly import data from compliant Web Feature Servers (WFS) (OGC, 2005) and export the results via a Web Map Server (WMS) (OGC, 2002) as a layer or in a variety of georeferenced imagery formats. Thus a user need only install the latest version of GeoServer and add the required jar files to have a fully functioning system on a computer with Java installed. Ideally a user will be able to configure their system to pull data from a remote server which is serving up population data from a central WFS (see figure 1). WPS allows for the user to specify the input as coming from a sub-process, so a user can construct a model to calculate a more complex expectancy or Population At Risk (PAR) estimate. However, if needed, it is simple for the user to add required PAR data layers to their own GeoServer instance.

To allow for the generation and comparison of results from different spatial clustering methods,

the system described has the following methods:

- The GAM/K system (Openshaw, 1996) which carries out an exhaustive search by applying a range of circles to the whole spatial area of the data set. While this method is sure to find a cluster (if one exists) it can be prohibitive to carry out this level of search on large data sets.
- The rare disease cluster detection method of Besag and Newell (1991) searches for clusters by examining circles centered on cases with a radius determined by the $k$ neighbouring cases. This reduces the number of circles to be examined but the nearest neighbour calculation can be time consuming with large data sets.
- The SatScan algorithm (Kulldorff, 1997) makes use of a scan statistic calculated for circles centered on each population point and extended to include up to half the total population at risk. The paper is unclear on the preferred method to expand the circle so we opt to extend the circle point by point though this leads to issues with nearest neighbour calculations again.
- A random circle method formalized by Fotheringham and Zhan (1996) allows a quick but non exhaustive scan of the data set. For very large datasets a user might choose to search quickly using random circles across the whole map and then apply one of the other methods in a smaller rectangle constrained to interesting areas.

## Conclusions

This paper describes a system developed for epidemiologists to use to search large databases to find clusters of rare diseases (such as Childhood Leukemia). The system is made available as open source software and is based on standards compliant OGC services. Providing the system as open source allows it to be verified as secure to work in networked environments with confidential data and it allows the academic rigour of the algorithmic implementations to be assessed. The system allows the user to pull in Census data from servers that serve it via a WFS. The system can be readily installed locally and on Grid and Cloud computing infrastructures. Results are made available to the user using the WMS standard which allows for them to be overlayed with other data which allows for further geographical exploration of the data which may help to explain spatial clustering in the incidence data.

## References


Besag, J. and Newell, J. (1991). The Detection of Clusters in Rare Diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1):143–155.

Fotheringham, A. S. and Zhan, F. B. (1996). A Comparison of Three Exploratory Methods for Cluster Detection in Spatial Point Patterns. *Geographical Analysis*, 28(3):200–218.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods*, 26(6):1481–1496.

OGC (2002). *Web Mapping Service Standard 1.1.1.* Number 01-068r3. Open Geospatial Consortium.

OGC (2005). *Web Feature Service Standard 1.1.0.* Number 04-094. Open Geospatial Consortium.

OGC (2007). *Web Processing Service Standard 1.0.0*. Number 05-007r7. Open Geospatial Consortium.

Olsen, S. F., Martuzzi, M., and Elliott, P. (1996). Cluster Analysis And Disease Mapping: Why, When, And How? A Step By Step Guide. *BMJ: British Medical Journal*, 313(7061).

Openshaw, S. (1996). *Methods for Investigating Localised Clustering of Disease*, chapter Using a geographical analysis machine to detect the presence of spatial clusters and the location of clusters in synthetic data, pages 68–87. Number 135. IARC Scientific Publication, Lyon, France.

Openshaw, S., Turton, I., Macgill, J., and Davy, J. (1999). Putting the Geographical Analysis Machine on the Internet. In Gittings, B., editor, *Innovations in GIS 6*, chapter 10, pages 121–132. Taylor and Francis, London.

Robertson, C. and Nelson, T. (2010). Review of software for space-time disease surveillance. *International Journal of Health Geographics*, 9:16+.

Turton, I. (2008). GeoTools. In Hall, B. G. and Leahy, M. G., editors, *Open Source Approaches in Spatial Data Handling (Advances in Geographic Information Science)*. Springer, 1st edition.