Turton I., Turner A.G.D. (2011) Putting the Geographical Analysis Machine on the Internet Revisited. Full paper prepared for the GeoComputation 2011 conference, London, 2011-07-19 to 2011-07-22.

Putting the Geographical Analysis Machine on the Internet Revisited

Ian Turton *, Andy Turner **

* Independent researcher ijturton@gmail.com
** CCG, School of Geography, University of Leeds, A.G.D.Turner@leeds.ac.uk

Metadata

- Draft version 0.3.x (2011-07-26) of a full paper prepared for the GeoComputation 2011 Conference which took place from 2011-07-19 to 2011-07-22 in London.
- Conference Abstract https://docs.google.com/document/d/1ehXcwFAz7Sictjb28LBA1Qx6Dk65FjK7aDrZ1qultjA/edit?hl=en_US

Contents

1. Introduction

On the 16th of December 2010 a Google code repository was set up to refactor some Java code for spatial cluster detection based on the Cluster tool source code developed in the SPIN!-project (Turner, 2011a). The repository is located at the following URL:

- https://code.google.com/p/spatial-cluster-detection/

First and foremost, the aim of this refactoring was to initialise an open development of space-time attribute pattern analyzers based on previous work we did, and to encourage others to get involved in developing and using the code in a coordinated fashion. This work is based on previous work led by our friend and one of the pioneers of this work in the 1980s and 1990s - Stan Openshaw. Sadly, Stan retired following a stroke over a decade ago (Wikipedia, 2011). If we did not resurrect this work, then we would be even more guilty of GIS crime (Openshaw, 1993).

As far as we are aware, no others have initiated an open development of space-time attribute pattern analyzers written in the Java programming language. The spatial cluster detection Java code developed in the SPIN!-project (which was made available as open source under the GNU Lesser General Public License (GNU, 2007) license in 2003) should have been committed to an open source code repository and its development encouraged long ago. However, the timing of this work is perhaps good as contemporary computational infrastructure now lends itself more readily to the task of space-time attribute data pattern detection. Also, Stan approaches the age of 65 years old this year which is currently the official retirement age of men in the UK. So the time is right for revisiting all his work - not least the work on geographical clustering.

The following is a definition of geographical clustering from the smart spatial analysis Geographical Analysis Machine (GAM) web pages:

"The simplest way of defining a cluster is as a localised excess incidence rate that is unusual in that there is more of some variable than might be expected. Examples would include: a local excess disease rate, a crime hot spot, a unemployment black spot, unusually high positive residuals from a model, the distribution of a plant or surging glaciers or earthquake epicentres, pattern of fraud etc. Virtually any variable that has a geographical distribution can be input into GAM. The assumption is that identifying these extreme areas (or outliers or unusual areas) may be useful in that there could be implicit geographical associations with other variables that can be identified and would be of interest. Pattern detection via the identification of clusters is a very simple and generic form of geographical analysis that has many applications in many different contexts. The emphasis is on localised clustering or patterning because this may well contain the most useful information." (Turton, 1998)

Section 2 provides some further background to this work. Section 3 focuses on motivation for this work. Section 4 provides implementation details. Section 5 presents some results with a very brief discussion. Section 6 concludes.

2. Background

The first on-line version of the GAM/K Geographical Analysis Machine was made available for others to use as part of project entitled "A smart spatial pattern explorer for the geographical analysis of GIS data" funded by the UK Economic and Social Research Council (Turton, 1998; Openshaw et al., 1999). The user of the system would upload input files and set parameters via some web forms and would then be notified when the results were ready to be downloaded. This on-line version of GAM/K (and a similar tool called the Geographical Explanation Machine or GEM) was available for about a year until catastrophic hardware failures occurred and the system was not recovered. The system failure coincided with an even bigger catastrophe - Stan Openshaw who had pioneered the development of GAM/K suffered a severely disabling stroke in 1999 and retired unable to continue working and with very diminished abilities to communicate.

Despite on-going struggles with health, Stan remains happy within himself and aware of his work and the world around him. Stan may one day surprise us all again and return to the stage to tell us about something new and exciting, but sadly this is very unlikely. To celebrate his work, in particular his contribution to GeoComputation, an event is tentatively scheduled to coincide with the GISRUK 2012 conference in April. So, this work is in part preparation for this celebration event, but more importantly, it is about developing and applying geographical analysis technology in an attempt to explore and discover problems and risk and thus help in alleviating them.

The first efforts to make a web service for GAM/K and the subsequent failure of the system provided a useful learning experience. At the same time, a first translation of the GAM/K code from the Fortran language into Java language was done. Back then, in the late 1990s, Java was in its infancy and without a core package for dealing with collections, but it was recognised as a language for the future - not least because Java programs were and are supported on a wide range of platforms and the language is designed to support protocols and communication between many devices on the Internet. Technology written in Java is therefore highly portable and capable of being made interoperable with other technology. Java programs can be readily designed to offer an uncomplicated push button solution that enables non-expert computer users to analyse data they have access to.

An extreme refactoring was undertaken for the SPIN!-project. Essentially, the Java source code was written again from scratch based on what was learned from the first attempt. Gaining skills, knowledge and experience, a new and improved implementation was developed. In addition to GAM/K, various other spatial cluster detection methods were implemented. The SPIN!-project was funded via the European Commission for 3 years and ran from the start of the year 2000, (Turner 2011b). Work package 5 was to build on the many years of work developing what were now collectively called geographical analysis machines (Openshaw 1995, 1999; Turton et al. 2000). A standalone tool called Cluster was developed that incorporated various geographical clustering methods including GAM/K. The Cluster tool was further integrated into a Spatial Data Mining System called the SPIN!-system (Turner 2011b).

A web based SPIN!-system (which incorporated Cluster) was made available at a SPIN!-project partner institution, but this system became unavailable just over a year after the end of the project. The Cluster component was stand alone, it and the source code were made available online at the end of the SPIN!-project (Turner 2011a). The Cluster source code and compiled program are still available and work with most if not all versions of the Java Virtual Machine (JVM).

Since the end of the SPIN!-project our attention has been diverted to other things. We had hoped to continue with this work, but were unsuccessful in attempts to get funding. Motivation for putting right the abandonment of this work was often sparked by user interest. In December 2010, the spark from one user re-ignited the fire and the code was developed on a whim rather than with any direct funding. It seemed timely as the NeISS project aimed to provide a version of GAM/K as a hosted service as part of its offering (Turner 2010).

3. 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, export 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 restrictions 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 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 secure in the system. Additionally, being open source allows for the academic rigour of the algorithm implementations to be assessed. Arguably there is too great a risk, with closed source system, regarding the security of confidential data.

4. Implementation

The initial aim was to re-implement the GAM/K part of the Cluster source code to make use of the GeoTools process API (Davis 2008), and the Open Geospatial Consortium (OGC) Web Processing Standard (WPS). The GeoTools process API was to be used as the main controller, and WPS was to be used to assist in loading data and delivering results.

An initial refactoring was done and the famous Gateshead Cancer Cluster result (from Openshaw et al. 1987) was reproduced. The original output is shown in Figure 1, the new output is shown in Figure 2. Google code project hosting was used as a repository for this refactored code.

The system makes use of the GeoServer WPS service implementation. GeoServer is 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 data layer or in a variety of georeferenced imagery formats.

A user need only install the latest version of GeoServer and add the required jar files to have a fully functioning system (on a networked computer with Java installed). Ideally the 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. 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. The user can also add data layers to their own GeoServer instance.

Having completed the initial aim of getting GAM/K working, other methods that were also implemented in Cluster tool were refactored. In all, the system has the following methods:

- GAM/K (Openshaw, 1996) which carries out an exhaustive search by applying a range of circles to the whole spatial area of the data set. The ranges of circles search across a range of scales as specified by the user. At each scale the level of overlap between the circles can also be set by the user.
- Kth nearest neighbour (Besag and Newell, 1991) which searches for clusters by examining circles centered on case points with a radius determined by the k neighbouring cases. Compared to GAM/K, this is a more focussed and less computationally demanding search
- SatScan (Kulldorff, 1997) which 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. In our implementation, the circle is extended point by point involving nearest neighbour calculations.
- Random (Fotheringham and Zhan, 1996) which allows a quick but non exhaustive scan of the data set.

5. Application

In January 2011 the UK Government released geocoded crime statistics for England and Wales and made them available at a high level of spatial resolution via the following URL:

- http://police.gov.uk

Soon after their release, maps of the concentration of crimes were made available via the following URL:

- http://www.gis-tech.co.uk/CrimeHeatMap/main.html

See for example Figure 3.

Such 'heat maps' give a picture of general activity, but they do not indicate where concentrations of crime are unusual.

Mapping crime rates is entirely more appropriate. A crime rate is the number of crimes of some type, normalised by something. The normaliser could be the number of dwellings for burglary crime, or residential population or daytime population. It could also be comparable numbers of the same type of crime from an earlier time period. This latter type of normalisation can help identify where crime rates are significantly changing over time both proportionally and absolutely. (An increase from 1 crime to 2 crimes is the same proportionally as an increase from 10 to 20, but in terms of crime counts, clearly the larger absolute increase is more important.

Figures 4, 5 and 6 provide GAM/K output for burglary, robbery and Anti-Social Behaviour Orders in England and Wales. These are illustrative that the refactoring of GAM/K is working. The results are also interesting in themselves.

6. Conclusion

This paper describes a system developed and under further development for spatial cluster detection. Epidemiologists might use it to search large databases to find clusters of rare diseases (such as Childhood Leukemia). Criminologist might use it to examine patterns of crime and the changes in these patterns over time. Road safety analysts might use it to examine road traffic accident incidence over time. There are many other uses for such a system. An implementation of GAM/K is provided and this can be used as a geographical model building tool to test for patterns in the residuals of a predictive model when compared to a sufficient sample of observations.

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. This facilitates further geographical exploration of the data which may help to suggest reasons that eventually lead to explanations of spatial clustering in the incidence data.

It had been hoped that by the time of the GeoComputation Conference that a hosted service based on the latest code refactoring would be available as part of the NeISS e-Infrastructure. However, this has not happened and NeISS has satisfied itself with hosting a GAM/K service based on the Cluster code from the old SPIN!-project which is still available on-line.

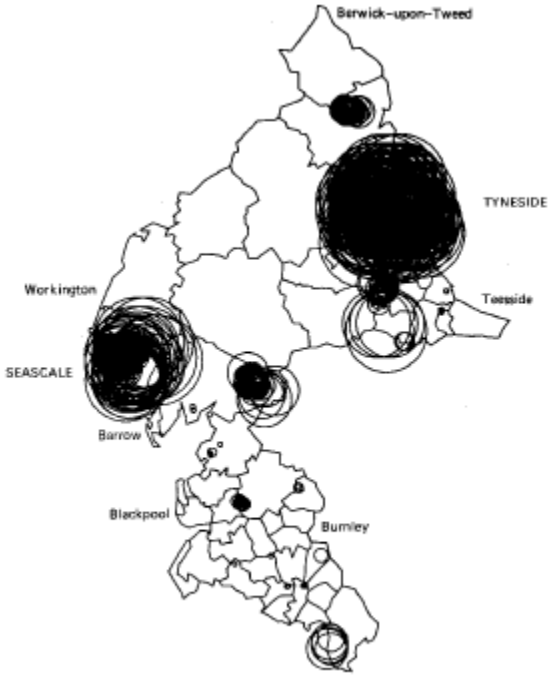
References/Bibliography

- Besag J., 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.
- Brunsdon C., Openshaw S., Macgill J.R., Turner A.G.D., Turton I. (1999) 'Testing space-time and more complex hyperspace geographical analysis tools'. Paper presented at the GISRUK conference, Southampton, UK, (1999-04). [online] http://www.geog.leeds.ac.uk/people/a.turner/publications/archive/BrunsdonEtAl1999.html [Accessed on 2011-07-25]
- Brunsdon C., Charlton M.E. (2010) An assessment of the effectiveness of multiple hypothesis testing for geographical anomaly detection (http://www.envplan.com/epb/fulltext/b38/b36093.pdf).
- Conley, J., Gahegan, M. and Macgill, J.R. (2005), A Genetic Approach to Detecting Clusters in Point Data Sets. Geographical Analysis, volume 37: pages 286-314. doi: 10.1111/j.1538-4632.2005.00617.x
- Conley J.F. (2004) A Genetic Approach to Automated Cluster Detection in Point Datasets. Abstract submitted to GIScience 2004. [online] http://www.geovista.psu.edu/publications/2004/Conley_GIScience2004_Abstract.pdf [Accessed on 2011-07-25].
- Davis G. (2008) GeoTools Users Guide: Implementing a new Process. [online] http://docs.codehaus.org/display/GEOTDOC/Implementing+a+new+Process [Accessed on 2011-01-04].
- Fotheringham A.S., Zhan F.B. (1996). A Comparison of Three Exploratory Methods for
- Cluster Detection in Spatial Point Patterns. Geographical Analysis, 28(3):200–218.
- GeoServer (2010) Geoserver Web Pages. [online] http://geoserver.org [Accessed on 2011-01-04].
- GeoTools (2010) GeoTools Web Pages. [online] http://www.geotools.org [Accessed on 2011-01-04].
- GNU (2007) GNU Lesser General Public License [online] http://www.gnu.org/licenses/lgpl.html [Accessed on 2011-08-01].
- 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., Elliott P. (1996) Cluster Analysis And Disease Mapping: Why, When, And How? A Step By Step Guide. BMJ: British Medical Journal, 313(7061).
- Open Geospatial Consortium, 2007, OGC Web Processing Service Standard. [online] http://www.opengeospatial.org/standards/wps [Accessed on 2011-01-04].
- …
- Openshaw S, Charlton M.E., Wymer C., Craft A. (1987) A Mark 1 Geographical Analysis Machine for the Automated Analysis of Point Data Sets. In the International Journal of Geographical Information Systems, Vol. 1, No. 4, pages 335-358. [online] http://www.informaworld.com/openurl?genre=article&issn=1365-8816&volume=1&issue=4&spage=335 [Accessed on 2011-01-04].
- Openshaw S., Turton I., Macgill J., Davy J. (1999) Putting the Geographical Analysis Machine on the Internet. In Gittings, B. (ed.), Innovations in GIS 6, chapter 10, pages 121-132. Taylor and Francis, London. [online] http://www.informaworld.com/openurl?

genre=article&isbn=978-0-7484-0886-3&volume=1&issue=4&spage=121 [Accessed on 2011-01-04].

- …
- Openshaw S. (1993) GIS 'crime' and GIS 'criminality'. Environment and Planning A, volume 25, number 4, pages 451-458.
- Openshaw S. (1995) Developing Automated and Smart Spatial Pattern Exploration Tools for Geographical Information Systems Applications. In the Journal of the Royal Statistical Society. Series D (The Statistician) Vol. 44, Number 1, pages 3-16. [online] http://www.jstor.org/stable/2348611 [Accessed on 2011-01-04].
- 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. (1999) Geographical data mining: key design issues. Paper presented at GeoComputation 1999. [online] http://www.geocomputation.org/1999/051/gc_051.htm [Accessed on 2011-07-25].
- …
- Robertson C., Nelson T. (2010) Review of software for space-time disease surveillance. International Journal of Health Geographics, 9:16. [online] http://www.ij-healthgeographics.com/content/9/1/16 [Accessed on 2011-07-22].
- …
- Turner A.G.D. (2010) Andy Turner's NeISS Project Web Page. [online] http://www.geog.leeds.ac.uk/personal/a.turner/projects/e-ISS/ [Accessed on 2011-01-04].
- Turner A.G.D. (2011a) Andy Turner's SPIN!-Project Web Page. [online] http://www.geog.leeds.ac.uk/people/a.turner/projects/SPIN/ [Accessed on 2011-07-25].
- Turner A.G.D. (2011b) CCG SPIN!-Project Web Page. [online] http://www.ccg.leeds.ac.uk/projects/spin/ [Accessed on 2011-01-04].
- …
- Turton I., Brunsdon C., Openshaw S., Macgill J.R., Turner A.G.D. (2000) Testing space-time and more complex hyperspace geographical analysis tools. In Atkinson P. M., Martin D. (eds.) GIS and Geocomputation (2000), pp. 87-102.
- …
- Turton I. (1998) Smart Spatial Analysis Web Pages [online] http://www.ccg.leeds.ac.uk/projects/smart/ [Accessed on 2011-01-04].
- Turton I. (2008) GeoTools. In Hall B.G., Leahy M.G. (eds.), Open Source Approaches in Spatial Data Handling (Advances in Geographic Information Science). Springer, 1st edition.
- ***...***
- Wikipedia (2011) Stan Openshaw Article [online] http://en.wikipedia.org/wiki/Stan_Openshaw [Accessed on 2011-08-01].

Figure 1. Acute Lymphoblastic Leukaemia clusters in Northern England (from Openshaw et al., 1987)



Significant circles at $p = 0{\cdot}002$ for acute lymphoblastic leukaemia.

Figure 2. The gateshead cluster result replicated in December 2010.

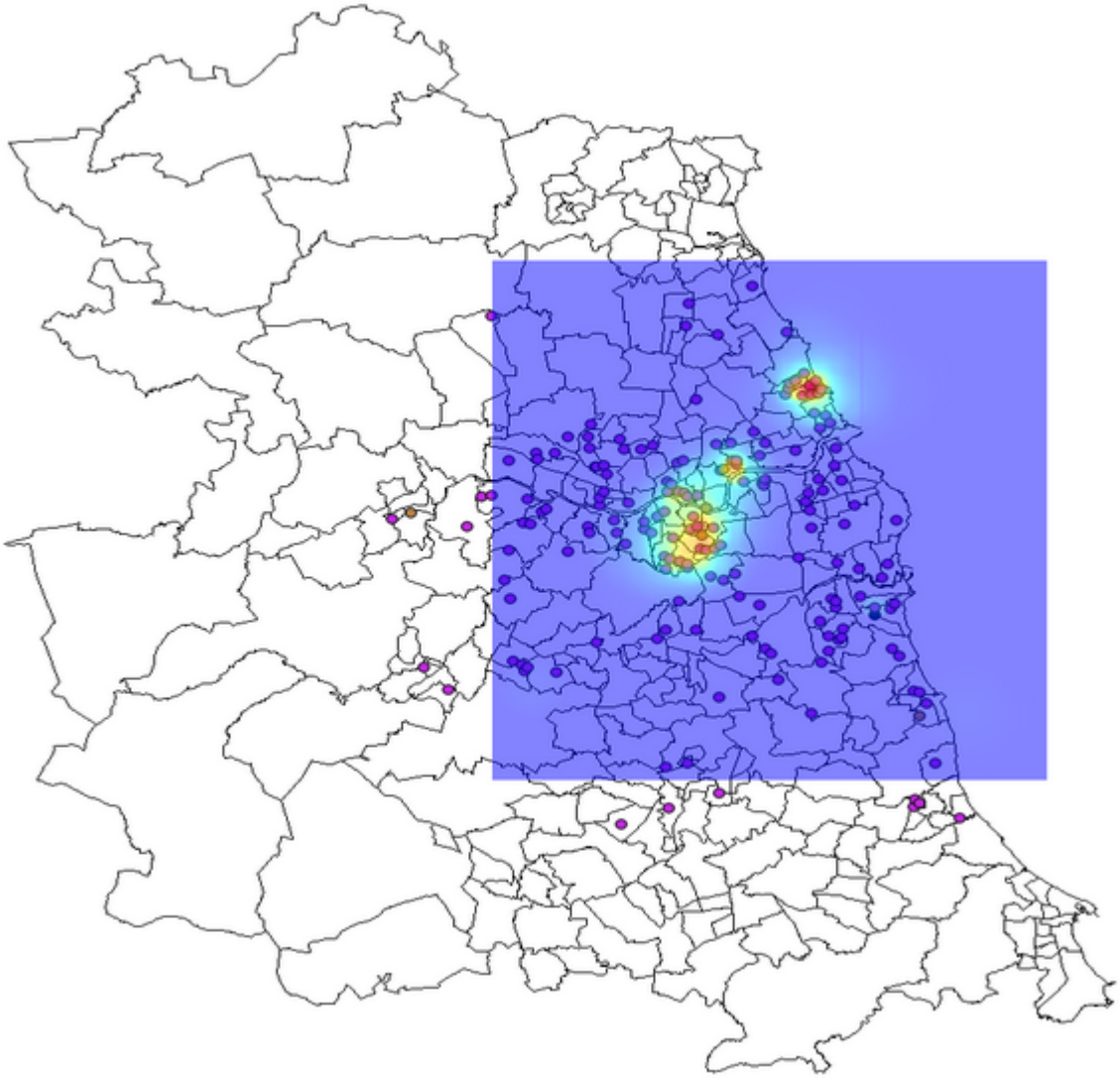
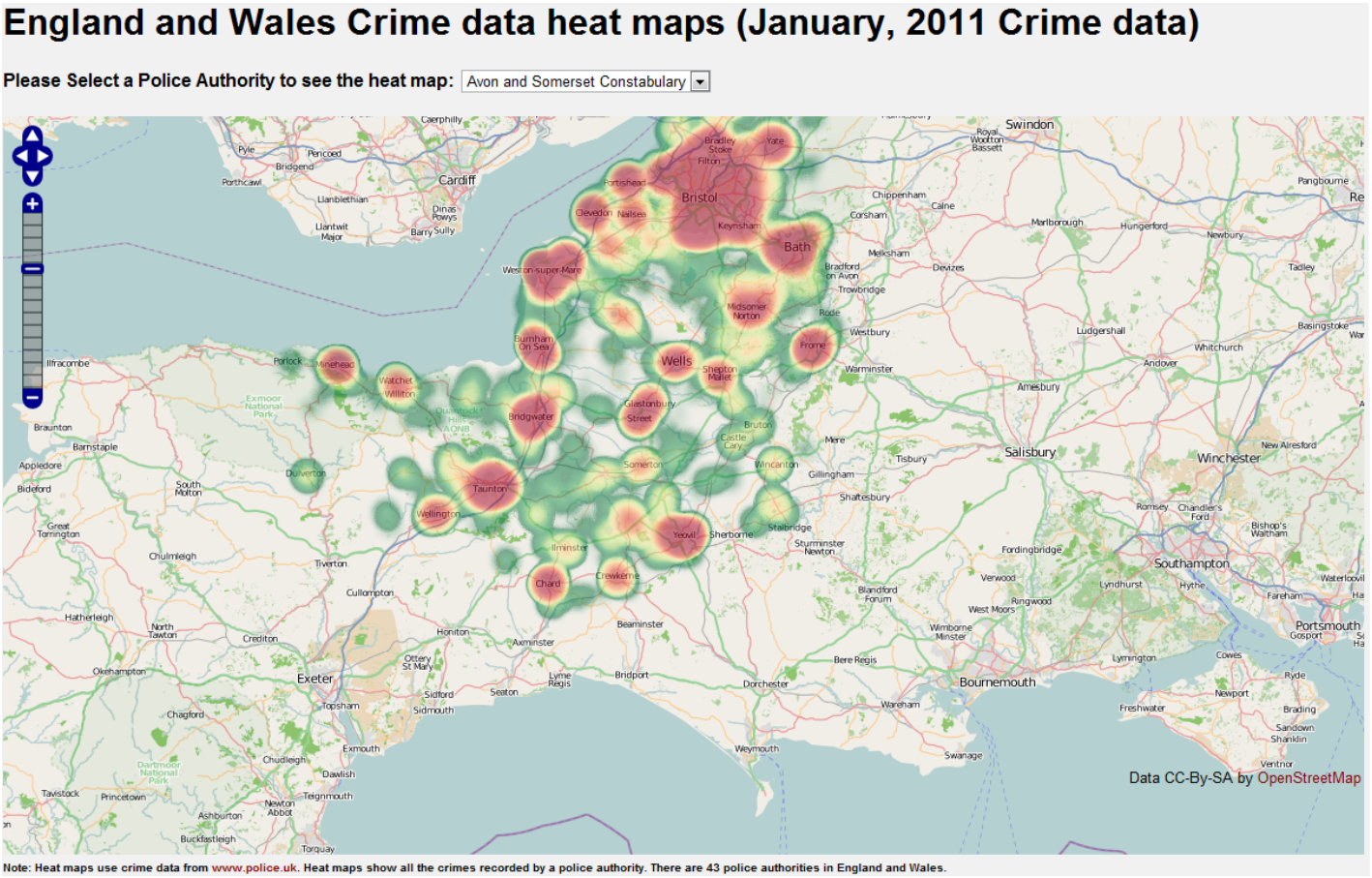


Figure 3. 'Crime Heat Map'



(source: http://www.gis-tech.co.uk/CrimeHeatMap/main.html)

Figure 4. GAM/K Burglary clustering (2011 crime data and 2001 census dwelling counts)

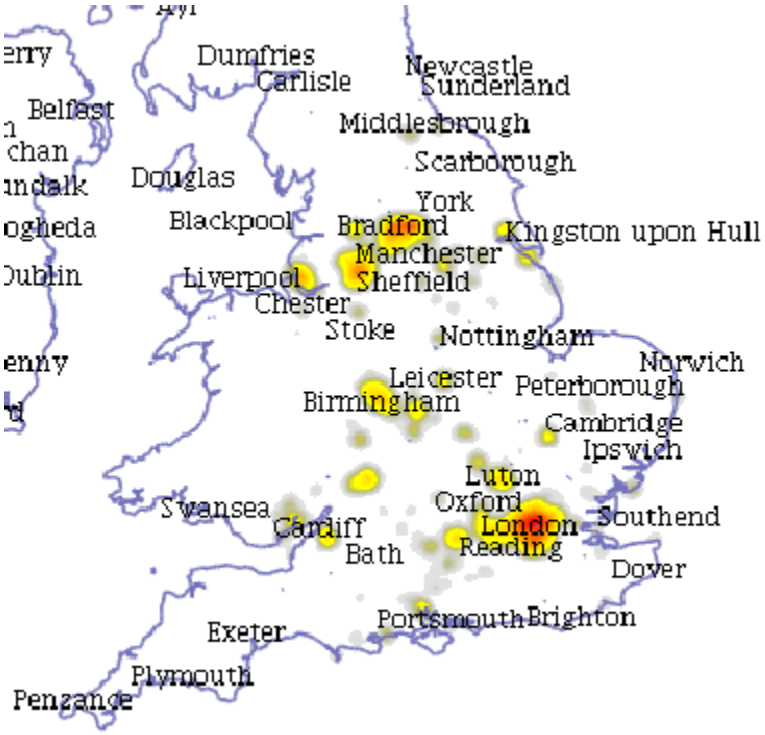


Figure 5. GAM/K Robbery clustering (2011 crime data and 2001 census residential population)

Figure 6. GAM/K ASBO clustering (2011 crime data and 2001 census residential population)