Spatially Clustered Associations in Health GIS "mashups"

Didier G. Leibovici[1], Lucy Bastin[2], Suchith Anand[1],
Jerry Swan[1], Gobe Hobona[1] and Mike Jackson[1]

[1]Centre for Geospatial Science, University of Nottingham,
Innovation Park, Triumph Road, NG7 2TU Nottingham, UK
Tel. +44 - (0)115 84 32760
didier.leibovici@nottingham.ac.uk, www.nottingham.ac.uk/cgs
[2]School of Engineering and Applied Science, Aston University, UK

KEYWORDS: spatial clustering, multivariate associations, co-occurrences, risk factors, Health GIS

1. Introduction

Developments in standards of the Open Geospatial Consortium (OGC) and International Organisation for Standardisation (ISO) along with the server-side and client-side software to allow the implementation of geospatial web services, enable GIS "*mashups*" to be seamlessly assembled by combining datasets from various sources and semantics. These geospatial "*mashups*" have huge potential in the health and epidemiological context, to derive intelligent outcomes, such as disease mapping or clustering, environmental risk factor analysis, exposure analysis, orforecasting and modelling of epidemics. However, practical application of these techniques requires efficient geoprocessing services that use pertinent statistical methods or algorithms; and there is frequently a dilemma in balancing the pertinence of the spatial methodology and the efficiency of the web service in terms of performance. This tradeoff between methodological complexity and real-time performance is amplified by the many and complex data sources which are available to be added to a '*mashup'*, and emphasises the need for simple exploratory methods which allow multivariate analysis of spatial data.

Cluster detection among labelled spatial features (in this instance, cases of disease) has a long history in epidemiology, ecology and geography (Lawson et al.,2006). Myriads of tests have been proposed to allow testing for spatial clustering or testing the *location* of clusters (e.g., Kulldorff , 2003, 2006). For locating of clusters, the Besag and Newell (1991) test, the GAM from Openshaw et al. (1987), and the spatial scan devised by Kulldorff and Nagarwalla (1995) are the most well-known approaches. While some principles of these methodologies differ (Besag and Newell is particularly distinct from the other two "scan" methodologies), they all make use of two populations: the cases, and the population at risk, (Waller and Gotway, 2004). For example, the spatial scan statistic implemented in SaTScan (Kulldorf 1997) allows use of a discrete Poisson model to handle 'case-only' data, but only in the presence of underlying population data for the region under consideration. Thus 'non-case' data is implicitly constructed to correct for any driving spatial heterogeneity in the population at risk.
With less control on the sampling design and/or the underlying population, disease mapping and disease clustering can be difficult, due to the heterogeneity of the overall background population and /or population at risk (*e.g.* if you are interested in mapping the incidence of a disease affecting only children under the age of five). Using the wrong population at risk, or working without 'non-case' data can easily result in misleading estimations of significance or pattern in a disease mapping application. Therefore "*mashups*" of case data and their potential risk factors need to focus on the more appropriate problem of locating spatial *associations*.
The aim of this paper is to present a generic approach for the challenge of detecting clusters of bivariate or multivariate associations between attributes of one or more populations or spatial features.

2. Multiway exploratory omnibus detection

Multinomial cluster detection can be seen as a simpler type of multivariate association cluster detection, if one considers each category as a realisation of a point process. This situation is shown in Figure 1, where categories do not seem to cluster themselves, apart from the **star** category which could show a cluster on the west "corner". Looking at the unmarked point pattern, one sees two or three clusters, but definitely *one sees only one cluster of association* of **star**, **dot** and **square** in the lower east "corner".

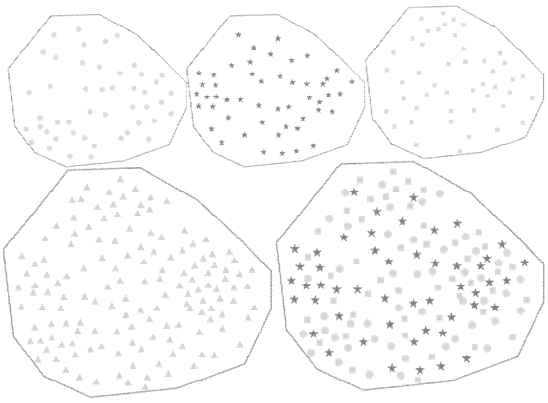

Figure 1. Hypothetical dataset of occurrences of three categories (top row): unmarked point pattern (bottom left) and marked point pattern (bottom right).

Recently Leibovici et al. (2008, 2009) developed an approach based on multiway contingency table co-occurrences of order ***k*** (***k>2***) to propose some exploratory methods allowing multinomial spatial dependence analysis. The CAkOO method uses a generalisation of correspondence analysis, (Leibovici, 2009), to decompose the chi-square of independence built from the multiway table whilst the SOOk method plots the entropy based on the multiway multinomial distribution of co-occurrences (for a chosen order) at different distances of collocation. CAkOO describes the spatial associations of categorical variables that are described without locating them, though some types of analysis allow spatial components to be displayed as well. SOOk, in a similar way to plotting a Ripley's ***K*** statistic, (Bivand et al.,2008), provides information on the spatial structuring of the co-occurrences at different scales (distances of co-occurrence).
These two methods are appropriate for overall detection of clustering / structuring by focusing on the problem of multiway and/or multivariate associations. However, they do not lend themselves directly to a delineated visualisation of association. Therefore, the declaration about Figure 1 -*one sees only one cluster of association*- needs a spatial scan approach to be fully assessed.

3. Clustered association detection,

The hypothetical data of Figure 1 could correspond to the identification of a multi-factorial zone of contagion (each point being a case and each category identifying a factor of contagion). Often the same point will carry more than one attribute (a multivariate point process) and multivariate multinomial co-occurrences analysis can identify profile clustering. The proposed method, called ScankOO, is inspired by the above-mentioned, widely-used cluster detection methods and exploits the spatial pattern that can be identified in high-order co-occurrences. The goal here is to build a statistical map that can be tested, using either Random Field theory or Monte Carlo simulation, for local maximum (or local minimum).

3.1 statistics for spatial association,

The two methods described above use two well known statistics describing associations: the chi-square of independence and the entropy, but are evaluated here on the contingencies of co-occurrences at a chosen order:

$$H_{Su}(C_{oo}, d) = -1/\log(N_{c_{oo}}) \sum_{c_{oo}}^{N_{c_{oo}}} p_{c_{oo}} \log(p_{c_{oo}}) \quad (1)$$

which defines a spatial entropy as the entropy, normalised to uniformity, of the multinomial distribution of co-occurrences at distance ***d***, with multi-index c_{oo} according to indexes of the categorical variables (attributes) and the co-occurrence order;

$$\frac{\chi^2}{N} = \sum_{ijk} p_{i..}p_{.j.}p_{..k}\left(\frac{p_{ijk} - p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}}\right)^2 \quad (2)$$

which is the chi-square of a co-occurrence table of order 3 for 3 variables indexed by ***i***, ***j***, and ***k.***. Different ways of computing co-occurrences were described in Leibovici et al.(2008) but the simplest to understand is that the maximum distance between the co-occurrent points (with defined labels) is at most ***d***.

3.2 scan of co-occurrences,

Different strategies of scan can be suggested. For example, for each point x_s of interest, a neighbourhood Vx_s is built to reach a condition of sufficiency (*e.g.* the number of points in Vx_s is exactly n_1, or the number of "cases" in Vx_s is n_2), then the statistics above (one or the other) are computed within the neighbourhood and attributed to the point x_s. The condition of sufficiency is fundamental here in order to ensure comparisons of the values obtained. Notice that ***d*** (distance of collocation) also plays an important role here; adaptation or optimisation with the parameters of the condition of sufficiency is an important aspect of the method (*e.g.* a range of ***d*** can be explored and part of the result is the maximum ***d*** at which the chosen statistic reaches a significance threshold). In a similar approach to GAM, one could also set the size and shape of neighbourhood, or as with a spatial scan, the neighbourhood could grow until the statistic is maximised (or minimised, in the case of spatial entropy).

4. Discussions

This methodology can be implemented within a parallel computing environment and distributed network using web services compliant to OGC standards. Specifically, the point data may be provided through Web Feature Services and the multivariate computation provided through Web Processing Services. The multivariate approach of ScankOO can allow *mashups* of several datasets, though care must be taken about sample size and consistency of sampling among data sources. There is a trade off between ***k,*** the order of co-occurrence, which mainly acts like a spatial constrainer, and the necessary sample size to build the multiway contingency table which estimates the multiway-multinomial distribution. Depending on the health or epidemiological study, the scan strategy - and particularly the condition of sufficiency for the neighbourhood - may be quite different. Example implementations onreal data from an MRSA study already used in Leibovici et al.(2008) will be presented at the conference.

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**, 143-155.

Bivand R.S Pebesma E.J and Gómez-Rubio V (2008) *Applied Spatial Data Analysis with R*.1st Edition., Springer-Verlag, New York Inc.

Ceyhan, E (2009) On the use of nearest neighbor contingency tables for testing spatial segregation. *Environmental and Ecological Statistics.* OnlineFirst

Kulldorff M and Nagarwalla N (1995) Spatial disease clusters: Detection and inference. *Statistics in Medicine*, **14(8)**, 799-81

Kulldorff M. (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods*,

26:1481-1496.

Kulldorff M Tango T and Park P.J (2003) Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, **42(4)**, 665-684.

Kulldorff M (2006) Tests of spatial randomness adjusted for an inhomogeneity: a general framework**.** *Journal of the American Statistical Association* 2006, **101(475),** 1289-305

Lawson A Gangnon R and Wartenberg D (2006) Developments in disease cluster detection. *Special Issue: Statistics in Medicine* **25, (5)**

Leibovici D.G Bastin L and Jackson M (2008) Discovering Spatially Multiway Collocations. *GISRUK Conference 2008, Manchester, UK, 2-4 April, 2008*

Leibovici D.G (2009) Spatio-temporal Multiway Decomposition using Principal Tensor Analysis on k-modes: the R package PTAk. *Journal of Statistical Software* (accepted August 2009)

Leibovici D.G Bastin L and Jackson M (2009) Higher Order Cooccurrences in Point Pattern Analysis and Decision Tree Clustering. *Computers & Geosciences,* (submitted)

Openshaw S Charlton M Wymer C and Craft A.W (1987) A mark I Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, **1**, 335-358

Waller L.A and Gotway C.A (2004) *Applied Spatial Statistics for Public Health Data*. Wiley, Hobooken, NJ.

Biography

Dr Didier Leibovici is a Research Fellow in geospatial modelling and analysis, with previous posts as a statistician in epidemiological/medical imaging research and as a geomatician for landscape changes in agro- ecology. Interests refer to interoperability and conflation models for cross-scales of integrated modelling applications within an interoperable framework chaining web services.

Dr Lucy Bastin is a Lecturer in GIS at Aston University. After a PhD on urban plant metapopulations, and research into fuzzy classification / uncertainty visualisation at Leicester University, she spent 3 years as a GIS software developer. Her current research interests include Web Processing Services for automatic interpolation, and spatial epidemiology.

Dr Suchith Anand is an Ordnance Survey Research Fellow at the Centre for Geospatial Science, University of Nottingham. His research interests are in open source GIS, automated map generalization, geohydroinformatics, mobileGIS, location based services, optimization techniques and asset management.

Dr Jerry Swan is a Research Fellow in Geocomputation. His approach includes optimization, graph theory, symbolic computation, knowledge representation, machine learning, semantics and ontologies. Jerry is particularly involved in the Persitent Test Bed project commissioned by OGC, AGILE and EUROSDR.

Dr Gobe Hobona is a Research Fellow in data, geoprocessing, workflow modelling standards and is consultant for the Open Geospatial Consortium (OGC) on the GIGAS project. Gobe's interests focus on management systems for OGC web services based on Grid computing and on cloud computing.

Pr. Mike Jackson is Director of the Centre for Geospatial Science. Prior to this he worked in industry at QinetiQ, Hutchison 3G, Laser Scan in various geospatial specialist and executive roles and in research for NERC. Mike is non-executive director of the Open Geospatial Consortium, and has research interests in combining new technologies such as positioning, pervasive computing and location based services for geo-informatics applications