

Cluster discovery in spatial data mining: a variable resolution approach

A. J. Brimicombe

Centre for Geo-Information Studies, University of East London, UK

Abstract

Spatial data mining seeks to discover meaningful patterns from data where a key dimension of the data is geographical location. This spatial dimension becomes important when data either refer to specific locations and/or have significant spatial dependence and which needs to be taken into consideration if meaningful patterns are to emerge. For point data there are two main groups of approaches. One stems from traditional statistical techniques such as k -means clustering in which every point is assigned to a spatial grouping and results in a spatial segmentation. The segmentation has k sub-regions, is usually space filling and non-overlapping (i.e. a tessellation) in which all points fall within a spatial segment. The difficulty with this approach is in defining k centroid locations at the outset of any data mining. The other broad approach searches for 'hotspots' which can be loosely defined as a localised excess of some incidence rate. In this approach not all points are necessarily assigned to clusters. It is the mainstay of those approaches which seek to identify any significantly elevated risk above what might be expected from an at-risk background population. Definition of the population at risk is clearly critical and in some data mining applications is not possible at the outset. This paper presents a novel variable resolution approach to cluster discovery which acts in the first instance to define spatial concentrations in the absence of population at risk. The cluster centroids are then used to establish initial centroids for techniques such as k -means clustering and arrive at a segmentation on the basis of point attributes. The variable resolution technique can thus be viewed as a bridge between the two broad approaches towards knowledge discovery in mining point data sets. The technique is equally applicable to the mining of business, crime, health and environmental data. A business-oriented case study is presented here.

1 Introduction

The 1990's were a period of transition for the spatial sciences from data-poverty to data-richness. Digital spatial data sets have grown rapidly in scope, coverage and volume [19]. This change has been facilitated by improvements in data collection technologies (e.g. GPS, remote sensing, point of sale), increasing computing power, falling costs of data storage (including the advent of data warehousing technologies) and the rise of Internet-based accessibility to data on-line. There has been an exponential rise in the size of databases, their structures have become commensurately more complex and the rate at which they can accumulate data is ever increasing. This has resulted in an urgent need for techniques that can mine very large databases for the nuggets of knowledge they contain. **Spatial data mining** can thus be defined as *techniques for meaningful patterns from large data sets where geographical location is a key interest*.

2 Discovering clusters in point event data sets

Geocoding of most databases in business, crime and health now takes place at the resolution of the individual address and can therefore be regarded as point event data sets. Whilst from the perspective of location each point can be treated as a binary occurrence, from a data perspective each point has added attribute dimensions further describing the location, the individuals involved or the nature of the event. Exploratory analysis of binary point events seeks to establish patterns from which causal spatial processes can be hypothesised or inferred [30]. Such analyses have had a long tradition in geography, ecology and epidemiology [5] [6] [14] [18]. In recent years, however, the adoption of a geocomputational approach to the analysis of spatial has represented a paradigm shift whereby computers are a pivotal ingredient of the science [9] [16] [23]. The tools for geocomputation naturally include geographical information systems (GIS) but are increasingly being used alongside neural networks, artificial intelligence, heuristics, spatial statistics, fuzzy computation, fractals, genetic algorithms, cellular automata and parallel computing [1]. Within quantitative geography there has also been a shift towards the local, that is, on exploring and understanding spatial differences between localities rather than quantifying their global similarities [8]. These have set the scene for renewed interest in analysing point event patterns [4].

The patterns detected in point event data are usually broadly classified as random, uniform or clustered. Generally, it is clustered patterns that produce the greatest interest as they most readily raise hypotheses of underlying spatial processes. Thus spatial cluster detection lies at the heart of spatial data mining [7] [11] [13] [19] [20] [21] [22]. Within the new context of geocomputation and local analyses, there are two broad approaches to cluster detection which have lead to a significant dichotomy in the meaning of 'cluster'.

The first set of approaches come from the mainstream statistics of cluster analysis arising from the work of Sokal and Sneath [26]. Thus clustering is an act of grouping by statistical means which, when applied to spatial data,

seeks to form a segmentation into regions or clusters which minimise within-cluster variation but maximise between-cluster variation. There is a general expectation that the spatial clustering mutually exclusively includes all points and is therefore space-filling within the geographical extent of the data. Examples of this approach are to be found in [12] [13] [20] [21]. A widely-used technique is the k -means clustering algorithm [17] which is relatively efficient in processing large sets of data. Weaknesses for data mining relate to its sensitivity to outliers [12] and that the number (k) of desired clusters or the location of k centroids needs to be specified from the beginning. This prior knowledge of k runs counter to the spirit of spatial data mining. Others [7] [11] have thus turned to Dirichlet and Delaunay diagrams respectively to define spatial clusters. These algorithms, however, will fail where points occupy the same location (as can often happen with postcode- or address-based geocoding) and to de-duplicate will inevitably lead to important data loss.

The second broad set of approaches is concerned with ‘hotspots’ as clusters. These can be loosely defined as a localised excess of some incidence rate and are typified by Openshaw’s Geographical Analysis Machine (GAM) and its later developments [22] [25]. Another approach is based around kernel density functions which map the highest densities as hotspots [10]. These hotspot approaches are central spatial epidemiology [15] in identifying from an at-risk background or control population areas of significantly elevated risk. Critical to these approaches is the definition of the at-risk population which in a data mining context may not be known with any precision at the outset. Mis-specification is clearly going to lead to erroneous results. Also, as a fundamental difference in the definition of ‘cluster’ from the first set of approaches where every point is assigned to a group, only some of the point events form hotspots and only these remain the focus of analysis. The approach outlined in the rest of this paper aims to bring the two approaches together.

3 Geo-ProZones: a variable resolution approach

Variable resolution is associated with the notion that scale and resolution are not treated as being uniform across an area but are allowed to vary spatially in response to a point event pattern. This is at the heart of the theory of adaptive recursive tessellations as first given in [27]. Spatial analytical applications are given in [28] with specific applications to point event patterns in [4]. The variable resolution approach to space is achieved through a recursive decomposition of space not dissimilar to quadrees but allowing for variable decomposition ratios and rectangular cells. No prior assumptions are made in the algorithm about the statistical or spatial distribution of point events. Each point is treated as a binary occurrence. The decision to further decompose any one cell larger than the atomic cell size is based on the variance that would occur at the next level of decomposition and a heuristic on the number of empty cells that result. The atomic cell size is mediated between the average nearest neighbour distance and average cell size per point. The algorithm has been shown, through tests, to be consistently effective in comparison with other approaches of

detecting point event clusters [4]. The resulting hotspots are termed Geo-ProZones (GPZ) since they represent geographical proximity zones in the point pattern. An example is given in § 4 below.

In the initial identification of GPZ clusters, an at-risk population is not used. This is because it may not be possible to do so at the outset in many data mining applications without risk of mis-specification. Proponents of GAM-type and pure epidemiological approaches would be dismissive of identifying clusters without reference to an at-risk or some control population. Point event (count) data on their own do nevertheless reflect workload, revenue stream or commitment of resources in meeting a spatially distributed demand. GPZ clusters provide just such a picture showing concentration of occurrences. Once that pattern and its attributes (including relevant the at-risk population) are clearly understood, then a second stage analysis of risk can be carried out.

4 Geo-ProZones as a precursor to *k*-means clustering

The proposal then, as illustrated below in a case study, is to use GPZ cluster centroids as a guide to setting up the *k*-means clustering. GPZ clusters are only concerned with the spatial distribution of binary events typical of a hotspot approach. The *k*-means clustering can then be used to introduce the other descriptive attributes of the points events to derive a spatial segmentation which incorporates all points. A further hotspot-type analysis can then be carried out within each segment in relation to the at-risk population appropriate to that segment in order to identify statistically significant clusters. This then brings together the two broad approaches to cluster detection in spatial data mining.

The case study is based on a one-year business transactions database. Customers were geocoded by address and the attribute SPEND (total spent by each customer in the year) was singled out for analysis. Through the postcode, each customer was assigned to one of ten higher order GB-Profiles lifestyle categories [24] given in Table 1. The business operates from a single hub in a provincial town in England and aims to serve a regional market. The spatial distribution, after removal of spatial outliers, is given in Figure 1.

Table 1: Higher order GB-Profiles lifestyle categories.

1	Struggling	Local Authority rented flats
2	Struggling	Local Authority rented terraced houses
3	Struggling	Local Authority rented semi-detached houses
4	Struggling	Privately renting terraces and bedsits
5	Aspiring	Privately rented bedsits and flats
6	Aspiring	Mixed tenure terraced houses
7	Climbing	Buying semi's and detached houses
8	Established	Owning or privately renting large detached houses
9	Prospering	Owning and buying large detached houses
10	Established	Owning and buying semi-detached houses

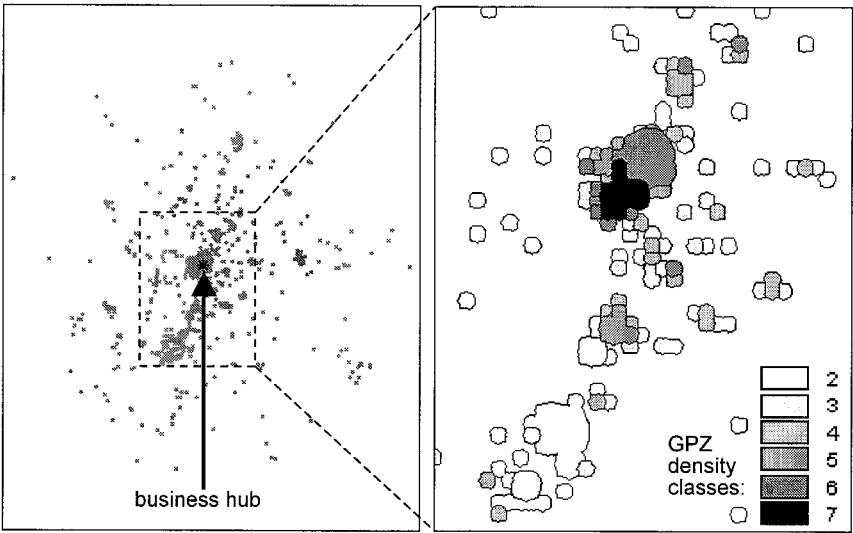


Figure 1: Distribution of customers. Figure 2: Pattern of GPZ clusters.

For databases which do not have any defined spatial boundaries (i.e. are not specifically linked to any boundary system), a convex hull is first fitted to the point events. GPZ clustering is then carried out. The result for the core area around the hub, where there are the highest densities, is given in Figures 2. The GPZ cluster pattern shows a trend from southeast to northwest through the business hub. On inspection of this pattern, eight cluster centres (mostly high density classes) were selected as the seed for the k -means clustering (Figure 3).

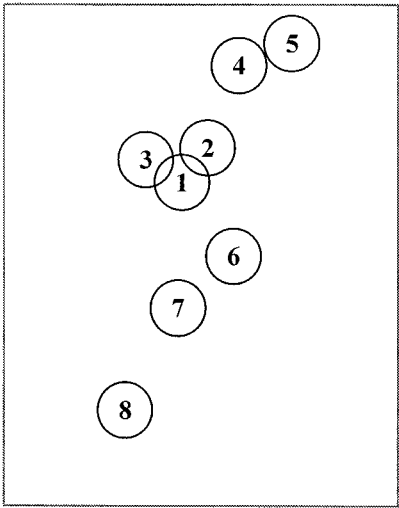


Figure 3: Location of $k = 8$ centroids in relation to Figure 2.

Distances from each of these centroids to all the customers were then calculated and entered as new fields. These distance fields and SPEND were then normalised using the technique of *robust normalisation* [2] [3]. Robust normalisation produces a distribution of median 0, lower quartile of -1 and upper quartile of +1 and is not sensitive to long tails as would a z-score normalisation (Figure 4). Robust normalisation also assists detection of extreme values, important in the *k*-means clustering.

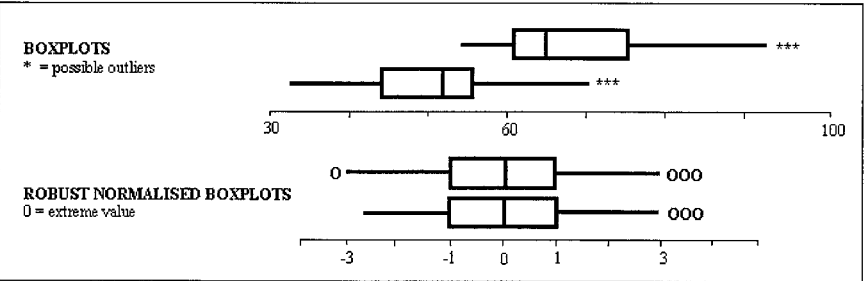


Figure 4: The principle of robust normalisation illustrated using boxplots.

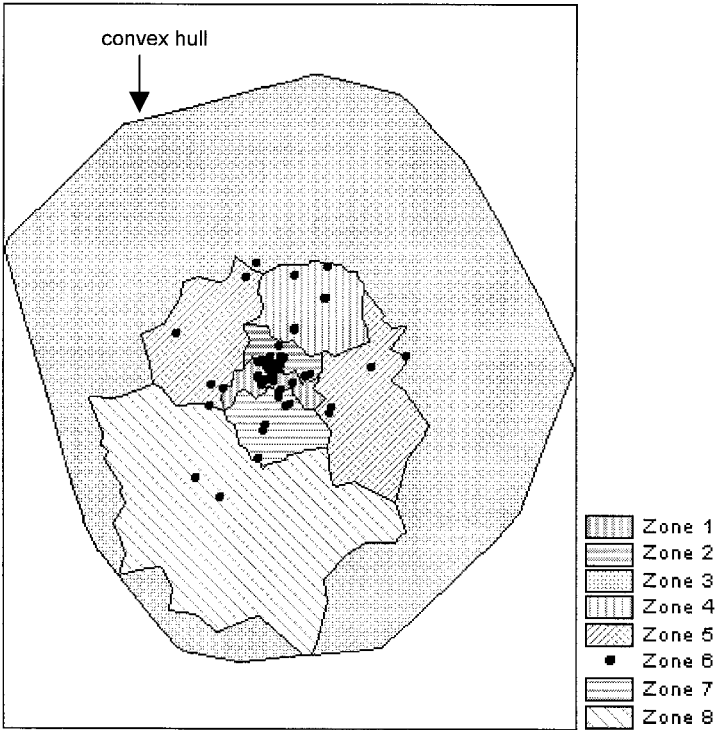


Figure 5: Result of *k*-means clustering using all attributes (same area Fig. 1).

The k -means clustering was run using the eight customers nearest to each of the eight GPZ cluster centroids and positioned as the first eight records in the data set so as to act as *a priori* cluster centres. The initial run of the k -means clustering included only the distance measures without other attributes. Since in this clustering the points were treated as binary events, the result was purely spatial and resulted in a mutually exclusive spatial segmentation into eight zones ($k = 8$). The pattern looked very similar to but not the same as Figure 5. Each cluster was distinct in terms of the mean scores for each of the data fields. GPZ centroids 4 and 5 (in Figure 3), however, were combined into a single zone thus allowing a zone to form around the periphery of the region which represents sparsely dispersed customers that are relatively far from the business hub. The analysis then progressed to a k -means clustering using all the attributes, that is, including SPEND and GB-Profile attributes. The result is shown in Figure 5. Table 2 summarises the characteristics of each of the resulting clusters. The k -means clustering is quite similar in overall appearance to the spatial only clustering but with the addition an unexpected result – one cluster (Zone 6 in Figure 5) that is not spatially mutually exclusive with the others. In order for this additional cluster to emerge, GPZ cluster centroids 6 and 7 (in Figure 3) were ‘merged’ to form Zone 7 in Figure 5.

Table 2: Characteristics of the $k = 8$ clusters (Zones 1-8 in Figure 5).

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8
n	602	469	178	185	107	70	218	195
Med. Dist	1164	1082	17100	5230	7228	1764	5929	13640
Tri. Spend	153	213	249	208	209	2884	226	247
GB-Prof1								2%
GB-Prof2			2%			1%		30%
GB-Prof3	8%	5%	4%	5%	2%	4%	14%	7%
GB-Prof5	2%	2%	3%				3%	3%
GB-Prof6	6%	4%	3%	6%		3%	5%	18%
GB-Prof7	47%	17%	24%	52%	7%	36%	9%	12%
GB-Prof8	7%	32%	25%	12%	36%	21%	28%	6%
GB-Prof9	15%	27%	28%	24%	52%	29%	35%	10%
GB-Prof10	15%	13%	10%		4%	6%	6%	12%

From inspection of Table 2, the eight k clusters (Z1 to Z8) have quite different characteristics, either spatially and/or in their attributes – median distance (in metres) from the business hub, trimean of SPEND and the GB-Profile classes. The trimean [29] gives a robust measure of central tendency that reflects the skewness of the distribution and is calculated as:

$$\text{trimean} = \frac{\text{lower quartile} + (\text{median} \times 2) + \text{upper quartile}}{4} \quad (1)$$

Although initially from Figure 2, GPZ cluster centroids 1 and 2 may be viewed as forming the same large cluster, their characteristics have emerged as being quite different. Z1 in Table 2 (Zone 1 in Figure 5) is dominated by geodemographic class 7 (*climbing*) and has the lowest trimean of SPEND both of which are in contrast to the neighbouring cluster Z2. Cluster Z5 (Zone 5 in Figure 5) is split spatially on either side of the centrally positioned clusters (Zones 1 and 2) immediately around the business hub. It nevertheless has a distinctive geo-demographic make-up being dominated by *prospering* and *established* households though not the highest spenders by any means. By contrast, Z8 which is comparatively far from the business hub and has 39% of its geodemographic profile in the three *struggling* classes is the third highest spending per customer. But probably of most interest from a business perspective is Z6 (Zone 6 in Figure 5), the cluster that spatially overlaps with the others. The customers from this cluster are the highest spenders, an order of magnitude above the others. They come from the *aspiring*, *established* and *prospering* geodemographic classes. Inspection of their distribution in Figure 5 suggests a spatial hotspot within this cluster that can be further analysed using GPZ clustering and tested for significance against the relevant background population. What is substantive is that such a small cluster (less than 5% of the customers) can be separated out from the larger database through the data mining process.

5 Conclusions

Demonstrated in this paper has been a dual approach to the spatial data mining of point event data. This a combination of hotspot-style clustering of the point events to identify k number of candidate centroids followed by a k -means clustering incorporating all attributes. The results of this dual approach assigns all points to clusters most of which are spatially mutually exclusive without precluding the emergence of distinct attribute-based clusters that overlap with the strongly spatial clusters. A follow-on stage, if necessary, is to repeat the hotspot-style clustering within each of the k -means clusters to establish those that are significant against the appropriate at-risk population. The hotspot-style clustering introduced in this paper has been a variable resolution approach which results in the identification of Geo-ProZones. It is the higher density classes of the GPZ that guide the choice of initial k centroids for the k -means clustering. Relevant variables for the k -means clustering were normalised using a novel technique of robust normalisation. Although the business example used here has been a relatively small data set, it has allowed the workings of the technique to be demonstrated in a succinct way. The variable resolution approach to producing GPZ clusters has shown itself to be effective as a means of guiding decisions on k where no other *a priori* knowledge is available. Significant for spatial data mining is that the overall dual approach can be used effectively on very large databases since no prior assumptions need be made about the content and structure of the point event data..

Acknowledgements

The author would like to thank Dr. Yang Li for his assistance in optimising the GPZ algorithms.

References

- [1] Armstrong, M.P. Geography and computational science. *Annals of the Association of American Geographers* **90**, pp. 146-156, 2000.
- [2] Brimicombe, A.J. Small may be beautiful - but is simple sufficient?" *Geographical and Environmental Modelling* **3**, pp. 9-33, 1999.
- [3] Brimicombe, A.J. Constructing and evaluating contextual indices using GIS: a case of primary school performance" *Environment & Planning A* **32**, pp. 1909-1933, 2000.
- [4] Brimicombe, A.J. & Tsui H.Y. A variable resolution, geocomputational approach to the analysis of point patterns. *Hydrological Processes* **14**, pp. 2143-2155, 2000.
- [5] Clark, P.J. & Evans, F.C. Distance to nearest neighbour as a measure of spatial relations in populations. *Ecology* **35**, pp. 445-453, 1954.
- [6] Cliff, A.D. & Ord, J.K. (1981) *Spatial Processes: Models and Applications*. Pion, London, 1981.
- [7] Estivill-Castro, V. & Lee, I. Argument free clustering for large spatial point-data sets via boundary extraction from Delaunay Diagrams. *Computers, Environment and Urban Systems* **26**, pp. 315-334, 2002.
- [8] Fotheringham, A.S. Trends in quantitative methods I: Stressing the local. *Progress in Human Geography* **21**, pp. 88-96, 1997.
- [9] Fotheringham, A.S. Trends in quantitative methods II: Stressing the computational. *Progress in Human Geography* **22**, pp. 283-292, 1998.
- [10] Gatrell, A.C.; Bailey, T.C.; Diggle, P.J. & Rowlingson, B.S. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers* **21**, pp. 256-274, 1996.
- [11] Halls, P.J.; Bulling, M.; White, P.C.L.; Garland, L. & Harris S. Dirichlet neighbours: revisiting Dirichlet tessellation for neighbourhood analysis. *Computers, Environment and Urban Systems* **25**, pp. 105-117, 2001.
- [12] Han J.; Kamber, M. & Tung, A. Spatial clustering methods in data mining. In Miller & Han (eds.) *Geographic Data Mining and Knowledge Discovery*. Taylor & Francis, London, pp. 188-217, 2001
- [13] Kiang, M.Y. Extending the Kohonen self-organizing map networks for clustering analysis. *Computational Statistics & Data Analysis* **38**, pp. 161-180, 2001.
- [14] Knox, E.G. The detection of space-time interactions. *Applied Statistics* **13**, pp. 25-29, 1964.
- [15] Lawson, A.B. *Statistical Methods in Spatial Epidemiology*. John Wiley & Sons, Chichester, 2001.
- [16] Longley, P.A.; Brooks, S.M.; McDonnell, R. & MacMillan, B. (eds). *Geocomputation: A Primer*. Wiley, Chichester, 1998.

634 *Data Mining III*

- [17] MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Maths and Statistics Problems* Vol 1 pp. 281-297, 1967.
- [18] Mantel, M. The detection of disease clustering and a generalised regression approach. *Cancer Research* **27**, pp. 209-220, 1967
- [19] Miller, H.J. & Han, J. (eds). *Geographic Data Mining and Knowledge Discovery*. Taylor & Francis, London, 2001.
- [20] Murray, A.T. Spatial characteristics and comparisons of interaction and median clustering models. *Geographical Analysis* **32**, pp. 1-18, 2000.
- [21] Murray, A.T. & Estivill-Castro, V. Cluster discovery techniques for exploratory spatial data analysis. *International Journal of Geographical Information Science* **12**, pp. 431-443, 1998.
- [22] Openshaw, S. Building automated geographical analysis and explanation machines. In Longley *et al.* (eds). *Geocomputation: A Primer*. Wiley, Chichester, pp. 95-115, 1998.
- [23] Openshaw, S. & Abrahart, R.J. (eds). *GeoComputation*. Taylor & Francis, London, 2000.
- [24] Openshaw, S. & Blake, M. *GB Profiler 91*. Department of Geography, University of Leeds, 1996.
- [25] Openshaw, S.; Charlton, M.E.; Wymer, C. and Craft, A.W. A mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* **1**, pp. 359-377, 1987.
- [26] Sokal, R. & Sneath, P. *Principles of Numerical Taxonomy*. Freeman, San Francisco, 1963.
- [27] Tsui, H.Y. & Brimicombe, A.J. Adaptive recursive tessellations (ART) for Geographical Information Systems. *International Journal of Geographical Information Science* **11**, pp. 247-263, 1997
- [28] Tsui, H. Y. & Brimicombe, A. J. Hierarchical tessellations model and its use in spatial analysis. *Transactions in GIS* **2**, pp. 267-279, 1997
- [29] Tukey, J.W. *Exploratory Data Analysis*. Addison-Wesley, Reading MA, 1977
- [30] Unwin, D. GIS, spatial analysis and spatial statistics. *Progress in Human Geography* **20**, pp. 540-551, 1996